DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

A CRITICAL LINK:

DIAGNOSIS TO PROGNOSIS

Proceedings of a Joint Conference

THE 51ST MEETING

of the

SOCIETY FOR MACHINERY FAILURE PREVENTION TECHNOLOGY

and

THE 12TH BIENNIAL CONFERENCE

on

RELIABILITY, STRESS ANALYSIS AND FAILURE PREVENTION
[RSAFP COMMITTEE OF ASME]

VIRGINIA BEACH, VIRGINIA
APRIL 14-18, 1997

Compiled by

HENRY C. PUSEY
and
SALLIE C. PUSEY

DTIC QUALITY INSPECTED 3

A Publication of the

Society for Machinery Failure Prevention Technology (MFPT)
(A Division of the Vibration Institute)

19970707 088

Copyright © 1997 by
Society for Machinery Failure Prevention Technology (MFPT)
(A Division of the Vibration Institute)
4193 Sudley Road
Haymarket, Virginia 20169-2420
All Rights Reserved

Special Notice

The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce, or allow others to publish or reproduce, the published forms of any papers in these proceedings authored by a government agency or a contractor to a government agency whenever such publication or reproduction is for U.S. government purposes.

TABLE OF CONTENTS

PREFACE ix

FOREWORD x

MFPT SOCIETY BOARD OF DIRECTORS xi

MFPT 30TH ANNIVERSARY xii

FEATURED PAPER

Symptom-Based Fault Diagnosis of Vibrating Structures 3
H. G. Natke

WORKSHOP ON MAINTENANCE ASSET MANAGEMENT

MIMOSA – A Key to Successful Information Age Equipment Asset Management 11
J. S. Mitchell

FAILURE ANALYSIS

Investigation of Main Rotor Spindle Assembly Retaining Rods for an Army Utility Helicopter Fabricated by Three Different Contractors 19
M. Pepi, V. Champagne, S. Grendahl and K. Bhansali

Dynamics of Layered Composite Beams and Plates with a Single Delamination 33
M. Krawczuk, W. Ostachowicz and A. Zak

Investigation of Elastic, Plastic, and Creep Response of a Polymeric Clip Using Experimental and Numerical Techniques 43
W. D. Bartlett and J. Sherwood

Effect of Intergranular Surface Attack on the Fatigue Properties of AM355 Material 53
V. Champagne, S. Grendahl, M. Pepi, K. Bhansali and G. Liu

Maintenance of Treadbelt Shoes for the Shuttle Transporter 67
R. Lee, P. Becker and N. Salvail

Take the Mystery Out of Your Motors 73
A. McMillan

DIAGNOSTICS AND PROGNOSTICS I

Turbomachinery Diagnosis and Prognosis Based on Actively Solicited Impulse Response 83
W. D. Marscher and C-W. Jen

Prognostic Issues for Rotorcraft Health and Usage Monitoring Systems 93
C. S. Byington, S. E. George and G. W. Nickerson

Hierarchical Failure Simulation for Machinery Prognostics 103
G. A. Lesieutre, L. Fang and U. Lee

DTIC QUALITY INSPECTED 3

Developing Prognostics for Diesel Engines 111
E. J. Lewandowski

Intelligent On-Line Detection of Cavitation in Hydraulic Pumps 121
J. P. H. Steele

Quantitative Determination of the Thickness Variation in Thin Shell Structures and Sensing of Failure Points 131
I. Onichtchenko, G. Dovgalenko, A. Kniazkov and G. Salamo

An Autonomous Diagnostic/Prognostic System for Shipboard Chilled Water Plants 139
T. G. Edwards and G. D. Hadden

DESIGN, FAILURE PREVENTION AND STRESS ANALYSIS

On the Redesign and Material Selection of a Pedal Spindle Used in Bicyles: A Case Study 151
M. Manoogian, D. Johansen, P. Crawford, R. Nishimuro, J. Foyos, F. Fisher and O. S. Es-Said

A Graphical Approach to Spring Design 161
M. L. Smith, F. E. Fisher and O. S. Es-Said

Infant Crib Failure Analysis Case Study 171
D. B. Brickman and R. L. Barnett

On the Corrosion of AISI 304 and Galvanized Steel at a Meat Packaging Company 181
D. E. French, I. Hernandez, R. W. Aston, J. M. Busse, M. L. Smith, J. Foyos and O. S. Es-Said

Stress Analysis of Weldments by Photostress 191
C-L. Chang and R-C. Lee

Interlayer Shear Slip Effects on the Load Transfer Mechanism in Nonsymmetrically Constrained Thick Laminated Strips 199
D. W. Swett

Effect of Steady-State Operating Temperature on Power Cycling Durability of Electronic Assemblies 219
J. M. Kallis and M. D. Norris

SENSORS TECHNOLOGY

Grating-Based Optical Fiber Sensors for Structural Analysis 231
V. Bhatia, V. Arya, T. D'Alberto, D. Sherr, D. Campbell, M. J. de Vries, R. O. Claus and C. P. Nemarich

Helicopter Gearbox Vibration Measurement with the Robust Laser Interferometer 241
H. M. Holland, T. C. Goodenow and R. L. Shipman

A Prognostic Monitor for Microelectronics Systems Reliability 251
W. W. Sifre, S. L. Drager and M. J. Walter

Highly Effective Real Time Adaptive Holographic Interferometry and Application to Nondestructive Testing in Machinery 261
G. E. Dovgalenko, A. V. Kniazkov, Y. I. Onischtchenko, G. J. Salamo

High Reliability Accelerometers for Use in Shipboard Condition Assessment Systems 271
R. M. Barrett, Jr.

Optical Fiber Sensors for Strain Measurement in High-Pressure Air Compressor Bolts 279
V. Bhatia, M. J. de Vries, K. A. Murphy, R. O. Claus and C. P. Nemarich

Real Time Interferometric Analysis of Strained Silicone Film 289
G. E. Dovgalenko, Y. I. Onischenko, A. V. Kniazkov, M. S. Haque, H. A. Naseem and G. J. Salamo

DIAGNOSTICS AND PROGNOSTICS II

AR Modeling with Dimension Reduction for Machinery Fault Classification 299
A. K. Garga, B. T. Elverson and D. C. Lang

Model-Based Tracking of Hidden Parameter Drift in Dynamical Systems 309
J. P. Cusumano, D. Chelidze and N. Hecht

The Algebraic Determination of Current Components in Induction Motor Driven Machinery 319
R. G. Mitchiner and M. Kaiser

Main Propulsion Reduction Gear Case Study 327
W. A. Watts

Optimizing Reliability, Maintainability and Cost for Shipboard Systems 337
J. A. Latimer, W. C. Momberger and N. Johnson, Jr.

Preemptive Maintenance and Intrinsic Health Monitoring 345
D. K. Hoth

PROBABILISTIC DESIGN, FAILURE ANALYSIS AND SAFETY METHODS

A G-Function-Based Probabilistic Design Methodology for System Reliability 359
W. Wang and Y-T. Wu

Correlation Procedures for Fatigue Life Determination 369
R. Kouta and D. Play

Comparative Statistical Analysis of Fatigue Crack Growth Data 379
J. D. Baldwin and C. A. Paul

Probabilistic Life Assessments of Arbitrary Shaped Scarred Boiler Tubes 389
K. Zarrabi and H. Zhang

Analysis of Electromechanical Systems Reliability Using Markov Models 399
J. Lu and J. D. Baldwin

A New Quantification of the Reliability of Components Subjected to All Types of Fatigue Loads 409
D. B. Kececioglu and W. Wang

Node-Based Structural Shape Optimization 419
X. Song and J. D. Baldwin

OIL ANALYSIS

Characterization of Debris from F404 Engine Oil Filters by Energy Dispersive X-ray Fluorescence 431
G. R. Humphrey

Real Time Optical Oil Debris Monitors 443
J. Reintjes, R. Mahon, M. D. Duncan, L. L. Tankersley, J. E. Tucker, A. Schultz, V. C. Chen, C. Lu, T. L. McClelland, P. L. Howard, S. Raghavan and C. L. Stevens

Filter Debris Analysis Using XRF 449
R. R. Whitlock

Condition Monitoring of Perfluoropolyether (PFPE) Lubricated Components 459
R. L. Silvestri

Rotrode Filter Spectroscopy Revisited - A Method for Multi-Elemental Analysis of Particles in Used Oil 471
D. P. Anderson and M. X. Lucas

Oil/Wear Particle Analysis a Predictive Maintenance Tool 477
R. Dalley

SIGNAL PROCESSING

Integrated Life Measurement and Diagnostics for Gas Turbine Engines 489
M. J. Roemer

Hybrid Reasoning Techniques for Automated Fault Classification 501
A. K. Garga and D. L. Hall

Gear and Bearing Diagnostics Using Neural Network-Based Amplitude and Phase Demodulation 511
E. C. Larson, D. P. Wipf and B. E. Parker, Jr.

Fault Severity Estimation Using a Neural Network Fault Tracking Approach 523
J. E. Lopez, K. Oliver and I. A. F. Yeldham

On-Line Wear Assessment of Milling Inserts Using Time-Frequency Distribution and Nonlinear Virtual Sensor 531
C. J. Li and T-C. Tzeng

Machinery Failure Diagnosis Using Discontiguous Detection Method 541
Y. Shao and K. Nezu

NUMERICAL AND ANALYTICAL METHODS

The Dynamics Stress Intensity Factor of Surface and Internal Crack Subjected to Obliquely Shear Waves 553
I-C. Weng and Z. Ji

Wavelet Transformation: Theory and Engineering Applications 563
D. Paul, S. M. Pandit and W. Hernandez

Failure Assessment of Cylindrical Structures Subjected to Internal Impulsive Spot Loading 573
M. Moatamedi, B. C. R. Ewan and J. L. Wearing

Analysis of Multiple Crack Problems in Finite Structures 579
M. R. Kratochwill, R. Greif and C. Mauge

Space Enrichment Method for Finite Element Approximation of Mixed Boundary Problem in Two-Dimensional Elasticity 589
Y. Volpert, B. A. Szabo

Dynamic Spherical Cavity Expansion of Thermal-Viscoplastic Material 599
Z. Jian and L. Chweeteck

DATA COLLECTION FOR VIBRATION DIAGNOSTICS

A Common Sense Approach to Data Collecting: Data Collecting is More Than Putting a Probe on a Point and Pushing a Button 609
J. Flowers

LIFE EXTENSION AND DURABILITY

Real-Time Reliability Evaluation of Vibrating Mechanical Structures 625
U. K. Rakowsky and D. Söffker

Monitoring Residual Stress in Machinery Using Holographic Techniques 637
G. Salamo, G. Dovgalenko, Y. Onischenko and A. Kniazkov

Measuring Residual Stresses Using a Small Blind-Hole and Phase-Shift Interferometry 641
A. Kniazkov, G. Dovgalenko, G. Salamo and T. Latishevskaia

Transitional Data for Estimation of Gearbox Remaining Useful Life 649
C. S. Byington and J. D. Kozlowski

Navy Submarine Machinery Vibration Monitoring Cost/Benefit and Performance Analysis 659
G. M. Milner and R. R. Smith, Jr.

Perspectives of Monitoring and Control of Vibrating Structures by Combining New Methods of Fault Detection with New Approaches of Reliability Engineering 671
D. Söffker and U. K. Rakowsky

Motor Systems Management 683
J. D. Kueck

The Value of Integrating Process Variable Monitoring with Vibration Monitoring - Total Condition Assessment 693
G. K. McGowan

FAULT DETECTION AND DIAGNOSTICS OF VIBRATING STRUCTURE - METHOD AND EXPERIMENTS

Model-Based Diagnosis - Methods and Experience 705
H. G. Natke and C. Cempel

Damage Detection of Dynamical Systems by SVD and Influence Coefficients 721
A. Lenzen

Dynamics of Multimodal Rotors with Transverse Cracks 731
P. Goldman, A. Muszynska and D. E. Bently

Case History of Failure Detection in the Dynamic Structure of a Factory Overspeed Test Facility Using "Moving Pictures" 741
E. M. Scheideler

Robust Fault Detection of Large Vibrating Structures by Means of Control Theory - Some Principal Remarks 751
D. Söffker

Bearing Fault Detection via High Frequency Resonance Technique with Adaptive Line Enhancer 763
Y. Li, J. Shiroishi, S. Danyluk, T. Kurfess and S. Y. Liang

Experimental and Numerical Study of Gear Faults in Gearbox: Evaluation of Dynamic Stress Fluctuations 773
J. Mahfoudh, C. Bard, C. Beigneux and D. Play

Fault Diagnosis of Rolling Element Bearings Using Artificial Neural Networks 783
L. Rajamani and R. Dattagupta

(Note: The paper above was presented in the Diagnostics & Prognostics I Session)

APPENDIX: MFPG /MFPT Publications 793

PREFACE

The 51st Meeting of the Society for Machinery Failure Prevention Technology (MFPT) and the 12th Biennial Conference on Reliability, Stress Analysis and Failure Prevention (RSAFP) were held jointly on April 14-18, 1997 at The Cavalier Hotel in Virginia Beach, Virginia. The MFPT Society is a Division of the Vibration Institute; the RSAFP Committee functions under the Design Engineering Division of the American Society of Mechanical Engineers (ASME).

The Technical Program was presented on Tuesday through Thursday, April 15-17, 1997. The Opening Session featured a Keynote Address by **Dr. J. Howard Maxwell** from Arizona's Palo Verde Nuclear Generating Station. His topic was *Diagnosis to Prognosis as a Part of a Comprehensive Predictive Maintenance Program.* Four other distinguished invited speakers included **Dr. Alfred Broz**, Federal Aviation Administration, on *Airworthiness Assurance Utilizing NDE for Civil Aviation*; **Mr. Scott M. Holliday**, National Aeronautics and Space Administration, on *A Corporate Approach to Institutionalizing Reliability Centered Maintenance*; **Dr. Erwin V. Zaretsky**, NASA Lewis Reswarch Center, on *Rolling Bearings - A Historical Perspective*; and **Mr. Raymond L. Thomas,** P.E., Union Camp Corporation, on *Application of Reliability Concepts in a Pulp and Paper Mill.* Plenary Lectures were presented on *The Design Process to Minimize Failure* by **Mr. Robert S. Ryan**, formerly of NASA's Marshall Space Flight Center; on the *Physics of Failure* by **Dr. William F. Braerman** of the University of Maryland; and on *Symptom-Based Fault Diagnosis of Vibrating Structures* by **Professor H. G. Natke** from the University of Wuppertal, Germany. Unfortunately, none of these excellent papers except an extended abstract of Professor Natke's Plenary Address is included in these proceedings.

Tabletop vendor exhibits were an integral part of the meeting, as well as a special commercially-oriented session where new products and services were described. In addition to the technical sessions, there were three panel/workshops. The first on *Maintenance Asset Management* was organized by **Mr. John S. Mitchell**. The second on *Data Collection for Vibration Diagnostics* was moderated by **Mr. James W. Taylor**. The third special session was organized by **Mr. Kenneth S. Jacobs**, Director of the Surface Ship Maintenance Division of the Naval Sea Systems Command. Highlighting the U. S. Navy's efforts to improve fleet maintenance, the session covered *Condition-Based Maintenance as Seen From the Viewpoint of Engineering Analysis, Technology and Customer Use.*

On behalf of the MFPT Society's Board of Directors, I extend thanks to our MFPT sponsors for supporting the conference. Kudos to Dr. Michael J. Roemer for doing a great job as the MFPT Society Program Chairman. We were pleased to have Dr. Shahriar Jahanian and the RSAFP Committee of ASME join us in organizing and presenting this successful event.

Henry C. Pusey
Executive Director
MFPT Society

FOREWORD

Structural reliability of engineered structures should be the paramount concern for design engineers. The increasing complexity and demands of today's structures makes it imperative that the design engineers have an understanding of all of the design factors that affect long-term reliability. Structural analysis is becoming more complex. The design engineer must now consider among other factors failure modes, reliability, safety techniques and probabilistic methods in their analysis. The 12th Biennial conference on Reliability, Stress Analysis and Failure Prevention; held April 14-18, 1997 in Virginia Beach, Virginia; was offered to address some of the concerns that design engineers face today. Most of the papers presented at this conference are contained in this volume. All of the papers presented cover the theory, design applications and case studies in the areas of stress analysis, reliability, and failure prevention. The sessions at the conference were entitled:

- Design, Failure Prevention and Stress Analysis
- Probabilistic Design, Failure Analysis and Safety Methods
- Numerical and Analytical Methods
- Fault Detection and Diagnostics of Vibrating Structure: Method and Experiments

As Chairperson for the 12th Biennial RSAFP Conference, I want to express my appreciation to all those who helped in making this conference possible - the authors, reviewers, session chairs and invited speakers. I would also like to thank Henry and Sallie Pusey of the MFPT Society for their cooperation and help in organizing and publishing the program. My thanks to Dirk Söffker for organizing a session for this conference and to Osama Jadaan, the previous conference Chair, for his valuable suggestions and help.

Shahriar Jahanian, Ph.D
Temple University

MFPT SOCIETY BOARD OF DIRECTORS

(The Society for Machinery Failure Prevention Technology is a Division of the Vibration Institute.)

OFFICERS - 1995-1997

Chair
MarjorieAnn E. Natishan, Ph.D.
University of Maryland
Mechanical Engineering Dept.
College Park, MD 20742

Vice Chair
John L. Frarey
JLF Analysis, Inc.
845 Worcester Drive
Schenectady, NY 12309

Secretary/Treasurer
Henry C. Pusey

Executive Director - Henry C. Pusey
Office Manager - Sallie C. Pusey
MFPT Society Headquarters
4193 Sudley Road
Haymarket, VA 20169-2420

STANDING COMMITTEE CHAIRS

Finance
Rudolph J. Scavuzzo, Ph.D.
University of Akron
College of Polymer Science, Bldg. 329
Akron, OH 44325-3909

Membership
Marc Pepi
U.S. Army Research Laboratory
AMSRL-MA-I, Bldg 400
APG, MD 21005-5059

Programs
G. William Nickerson
Pennsylvania State University/APL
PO Box 30
State College, PA 16804

Research
Kam W. Ng, Ph.D.
Office of Naval Research
Code 334
800 N. Quincy Street
Arlington, VA 22217

TECHNICAL COMMITTEE CHAIRS

Diagnostics & Prognostics
Howard A. Gaberson, Ph.D.
NAVFAC Engrg Service Center
1100 23rd Avenue
Port Hueneme, CA 93043-4370

Failure Analysis
Victor K. Champagne
U.S. Army Research Laboratory
AMSRL-MA-I Bldg. 400
APG, MD 21005-5059

Life Extension & Durability
Michael J. Roemer, Ph.D.
Stress Technology Inc.
1800 Brighton-Henrietta Town Line Road
Rochester, NY 14623

Sensors Technology
Henry R. Hegner
ManTech Advanced Technology Systems
6400 Goldsboro Road, Suite 200
Bethesda, MD 20817

Education & Training
James W. Taylor
Machinery Management Specialist
1018 Cherry Lane
W. Lafayette, IN 47906

Members-at-Large
Christopher Nemarich
NSWC/Carderock
Code 2758 3 Leggett Circle
Annapolis, MD 21402-5067

Paul L. Howard
Paul L. Howard Enterprises
1212 Clearbrook Road
West Chester, PA 19830

Mark T. Kirk, Ph.D.
Edison Welding Institute
1100 Kinnear Road
Columbus, OH 43212

THE MFPT SOCIETY CELEBRATES ITS 30TH ANNIVERSARY

On April 18-19, 1967, the organizing meeting for the Mechanical Failures Prevention Group (MFPG) was held at the Office of Naval Research. In celebration of that event, the names of the participants in that meeting are given below. Two of these gentlemen, **John L. Frarey**, the first FELLOW of the Society(1996), and **Erwin Zaretsky** were participants in this 51st MFPT Society Meeting.

Aerni, W. F.	- NAVSEC, Washington
Beary, Jr. A. D.	- Naval Ship R&D Center, Annapolis
Belt, J. R.	- NSRDC, Annapolis
Bocchi, W. J.	- RADC, Griffiss AFB
Bosco, R. J.	- NAVSEC, Washington
Brown, R. R.	- NAVAIR
Burch, J. L.	- NASA, MSFC
Byrnside, CAPT.	- Welcomed group
Chaikin, A.	- NAVSHIPS
Carlton, C. A.	- NAVSEC, Philadelphia
Doroff, S. W.	- ONR
DeMarco, R. J.	- NAVSEC, Washington
Fischer, J.	- ONR
Ford, S. H.	- NAVSEC, Washington
Frarey, J.	- Curtiss-Wright
Gross, W. A.	- Ampex
Grossman, N.	- AEC
Harned, A. M.	- Naval Materiel Command
Katcher, A. A.	- NAVSEC, Washington
Korbs, D. E.	- NAVAIR
Lewis, H. R.	- NAVAIR
Lichtman, E. A.	- NAVAIR
Lucca, L. L.	- NAVSEC, Washington
Lindstrorm, W. C.	- NAVSEC, Washington
Miska, J. K.	- NAVSEC, Washington
Orcutt, F. K.	- MTI
Peterson, R.	- NAVSHIPS
Robinson, LCDR.	- ONR
Ruffini, H. A.	- NAVSHIPS
Sawyer, J.	- Naval Materiel Command
Schrader, A. R.	- NSRDC, Annapolis
Smith, W.	- MEL, Annapolis
Stanley, R.	- NAVSEC, Washington
Sternlicht, B.	- MTI
Sollecito, W. E.	- General Electric
Tallian, Tibor	- SKF Industries
Tepper, H.	- Naval Materiel Command
Thomas, C. I.	- NAVAIR
Van Gelder, F. M.	- NATTS, Trenton, N. J.
Wagman, E.	- Naval Materiel Command
Whalen, D. L.	- NSRDC, Annapolis
Zaretsky, E. V.	- NASA, Lewis

FEATURED PAPER

Plenary Address

SYMPTOM-BASED FAULT DIAGNOSIS OF VIBRATING STRUCTURES

H. Günther Natke
Curt-Risch-Institute for Dynamics,
Acoustics, and Measurements, University of Hannover
Appelstr. 9 A
D-30167 Hannover, Germany

1 Extended Abstract: Conventional fault diagnostic methods for vibrating mechanical systems are well-known and discussed elsewhere, for example in the books [1-3]. Additionally, the many proceedings of the related conferences must be mentioned, such [4-8] which show the variety of themes and topics. Of the workshops held on the subject, only [9] will be mentioned. The newest book on the subject of diagnosis was published in 1996 by the author of this lecture, co-authored by C. Cempel, Poznań, Poland. It contains mainly the model-based diagnosis, but also summarizes the *classic* diagnostic methods [10]. The model-based method uses an adjusted mathematical model, adjusted to the condition states of the various requested life times of the system, which is also verified, validated and usable. A model of this kind is the best available knowledge base of the system. This subject is discussed in another paper of the author within this conference [11]. Here, conventional diagnostic methods are discussed (whatever the word conventional means) and it is unavoidable that these methods are subjectively chosen, as they reflect the interests of the author and his areas of expertise. Figure 1 shows both diagnostic procedures in principle.

The **system** here is the description of the real existing object in its environment: a bridge or an aircraft or a robotic drive mechanism. A **model** is a (homomorphic) map of the system, it is a simplified mathematical description of interesting aspects of the system: goal-orientated model (see the paper by the author in [12]). From a diagnostic point of view some features are usually not significant.

Models are uncertain; measured signals are uncertain, too. The types of uncertainties are generally different. But in order to reduce the **uncertainties**, we have to model them by the use of some information. These uncertainties, their models and their measures are described in [12]. Models of the system dynamics and uncertainty models of faulty systems and of the fault process itself (damage evolution) are particularly important in the design and optimization of diagnostic procedures.

A **fault** or **damage** will be defined as life time-dependent changes in the physical parameters of the model of the system under inspection. These are changes in the mass distribution, stiffness distribution, or/and in the damping distribution. They also may be expressed by changes in the eigenquantities. Two time variables have to be distinguished due to their different scales: these are the slow time coordinate of the system's life, and the fast time coordinate of the system dynamics. It is assumed

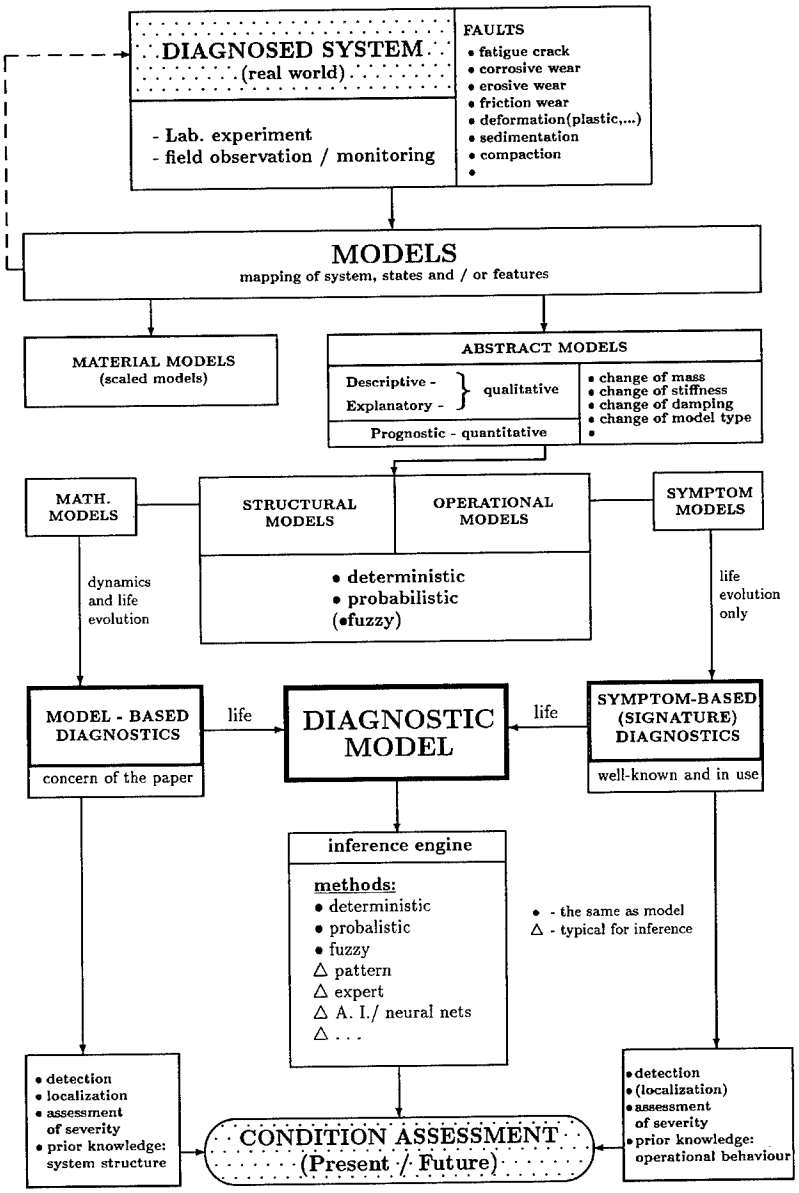


Figure 1: Flow-chart of the model-based and symptom-based diagnostic methods

that the physical properties of the system remain unchanged within a pre-defined life time interval. This assumption implies that no abrupt changes will occur during the inspection time (otherwise see [3]).

It is stated that the diagnosis of a fault/damage must result in detectable modifications of measurable quantities: **symptoms**. The distinguishability of faults by their symptoms or features plays a central role in diagnosis. The ability to distinguish between faults, or classes of faults, is a central criterion in evaluating both a diagnostic procedure and the models upon which it rests. Symptoms should have the following properties:

- directly or indirectly measurable
- functional relationship to a damage measure
- high sensitivity to a fault/damage as a local property
- distinguishability of various terms in the model, which also includes fault separation
- the absolute value is a non-decreasing function of time, unles the system is repaired etc.
- permit trend prediction.

Scalar and vectorial symptoms can be distinguished, as well as global and local symptoms. Scalar symptoms will lead to simple **signatures**, and vectorial symptoms to patterns that are classified as extended signature analysis. Symptoms of linearly behaving systems are different from those behaving nonlinearly. Symptoms can be assembled in a **feature vector** which can serve for decision-making. The features can belong to classes and form a **pattern**. Patterns have to fulfil some postulates which are presented in the lecture. Examples of symptoms, features and patterns will illustrate these explanations.

A description of some diagnostic methods and their assessment corresponding to that published in [13] will be given.

Various **damage measures** exist. Starting with the energy model of system damage, the relation between damage and symptom evolution is discussed. Another damage measure is based directly on the physical parameter modifications. The **damage capacity** and the **symptom-based reliability** are then introduced instead of the time-dependent reliability, and their diagnostic use is mentioned [14]. Different measures of system reliabilty, such as the measure of ability to perform a specific task, are taken into consideration. It has been shown that the commonly used life time-based reliability is not orientated towards application in the machine condition monitoring of critical systems. In this case it is much better to define and use the symptom-based reliability which allows one to measure directly the damage capacity, or **residual life**, and its future condition evolution.

Symptoms of dynamically behaving systems are time-dependent. Diagnostic experiments are generally based on a sample consisting of several units in order to perform satisfactory statistics. There are cases, for example especially in civil engineering,

where only one element/unit exists. The latter case means that one symptom or various symptoms of one unit are available. With the application of the **ergodic theorem** it can then be shown under particular conditions that both investigations lead to the same results, taking into account possible realizations of the **logistic vector**. The latter describes the system history.

Holistic modelling is the tool for diagnosis. This means taking the life time-dependent system modifications into consideration. The related mathematical model is therefore a time-variant model. This is generally handled, as already described above, by consideration of the different time scales as interval-wise time-invariant mathematical models. Instead of this procedure the **Hilbert transform** can be applied to the time-variant model, which together with the modulation theorem leads to an analytic equation (up to now of a single degree of freedom model) describing the holistic dynamics. The modulation theorem requires non-overlapping spectra of the related slow time varying parameters and the fast time dynamics, which can generally be provided. The equations for free and forced vibrations are presented [15].

In conclusion it can be stated that symptoms are very important in conventional as well as in *modern* diagnosis. New research results based on symptoms as classical quantities in diagnosis will underline this statement.

2 References:

1. Lyon, R.H., 1987, *Machinery Noise and Diagnostics*, Butterworths, Boston, London, Durban, Singapore, Sydney, Toronto, Wellington
2. Cempel, C., 1991, *Vibroacoustic Condition Monitoring*, Ellis Horwood, Chichester, UK
3. Basseville, M., Nikiforov, I.V., 1993, *Detection of Abrupt Changes: Theory and Application*, Prentice Hall Englewood Cliffs, New Jersey
4. Stanley, P. (ed.), 1992, *Structural Integrity Assessment*, Elsevier Appl. Scie. London, New York
5. Morgan, J., 1992, *Structures Congress '92, Compact Papers*, ASCE, New York
6. -, 1993, *Remaining Structural Capacity*, IABSE Colloquium 1993, IABSE Report, Vol. 67
7. Natke, H.G., Tönshoff, Meltzer, G., 1993, Dynamische Probleme - Modellierung und Wirklichkeit, Mitteilung CRI-K 3/93 des Curt-Risch-Instituts der Universität Hannover
8. Ruokonen, T. (ed.), 1994, *Fault Detection, Supervision and Safety for Technical Processes*, SAFEPROCESS '94, IFAC, Helsinki, Finland
9. Natke, H.G., Tomlinson, G.R., Yao, J.T.P. (eds.), 1993, *Safety Evaluation Based on Identification Approaches Related to Time-Variant and Nonlinear Structures*; Vieweg, Braunschweig, Wiesbaden
10. Natke, H.G., Cempel, C., 1996, *Model-Aided Diagnosis of Mechanical Systems - Fundamentals, Detection, Localization, and Assessment*, Springer-Verlag Berlin
11. Natke, H.G., 1997, Model-based diagnostics - Methods and experience, MFPT 51 & RSAFP 12, Proceedings

12. Natke, H.G., Ben-Haim, Y. (eds.), 1996, *Uncertainty: Models and Measures*, Akademie-Verlag Berlin
13. Ben-Haim, Y., Cempel, C., Natke, H.G., Yao, J.T.P., 1993, Evaluation of diagnostic methods, in: [9]
14. Cempel, C., Natke, H.G., Yao, J.T.P., 1995, Damage capacity and symptom-based reliability as advanced signal processing procedures, Int. Workshop on Structural Damage Assessment Using Advanced Signal Processing Procedures, Pescara, Italy, May 29-31, 1995
15. Natke, H.G., Cempel, C., 1995, System analysis and identification - A holistic approach, in: *Summer School on Systems Engineering - Collection of Lecture Notes*, Eds. H. G. Natke and C. Cempel, Mitteilung des Curt- Risch-Instituts der Universität Hannover, CRI-K1/95

MFPT

WORKSHOP ON MAINTENANCE ASSET MANAGEMENT

Chair: **John S. Mitchell**
Consultant

MIMOSA -- A KEY TO SUCCESSFUL INFORMATION AGE EQUIPMENT ASSET MANAGEMENT

by: John S. Mitchell
31882 Paseo Alto Plano, San Juan Capistrano, CA 92675

Abstract: Over the past twenty years, maintenance professionals have developed a rich array of highly successful practice and supporting technology to promote reliability, reduce cost and eliminate catastrophic failures. Planned Maintenance (PM) and Condition Directed or Predictive Maintenance (PdM) are examples of successful, broad use, proven practice. Condition assessment, lifetime prediction and diagnostic analysis based on vibration characteristics, operating fluid (lubricating and hydraulic oil) quality and condition, operating and electrical performance and thermography provide vital support.

Today, there is growing recognition that single technology solutions, no matter how well implemented, are insufficient to produce results required to meet the demands of modern, high intensity, operations. There is a clear need for an overall, guiding concept. The concept must incorporate current "best practice" maintenance and fully integrate all elements of equipment information. Information exchange with unit management and information systems is a necessity. Likewise, a compelling what and why to justify "how to" practice and technology is essential.

Many progressive visionaries are introducing a concept that some are beginning to call Asset Management. As typically presented, Asset Management is a strategic concept that goes beyond maintenance to include critical factors that influence performance, fix cost and ultimately determine the prime objective -- availability. These include specification, design, installation, safety, operation, economics, organization, maintainability, work practices and training. Information, or more specifically the integration, exchange and display of information is a key component.

Current and projected condition, supplied by multiple condition assessment systems and operating performance, conveys vital information. Design specifications, manufacturers' manuals, maintenance history, troubleshooting and task instructions, operating and maintenance recommendations (decision support) and lifecycle costs are other examples of crucial information. All must be immediately available and displayed intuitively for a broad range of users.

All have observed the unification, information access and intuitive display achieved by Microsoft Windows. The Internet is another example. The same principles must be extended to equipment asset management. The only question is how.

There is a solution. Demand open, self integrating, information systems. Vibration, lubricating oil, performance and electrical analysis systems that can freely exchange information between themselves, maintenance management and control systems. Think this is impossible? Personal computers seem to do very well with office applications.

MIMOSA, The Machinery Information Management Open Systems Alliance, has demonstrated feasibility with a multi supplier open exchange of vibration data. Suppliers are committed. All that is needed now is user demand to complete the vital task.

This presentation will summarize the necessity for and elements of change required to optimize the performance and availability of equipment assets. The advantages of fully integrated information as a fundamental element of success will be explored in depth. Details and current status of MIMOSA, an industry effort to gain open information exchange, will be presented.

Key Words: Asset management; condition assessment; condition directed maintenance; condition monitoring; maintenance; MIMOSA; predictive maintenance; vibration analysis

HISTORY -- IF YOU'RE NOT A STUDENT THE PAST IS BOUND TO REPEAT: Since the 1960's there has been general agreement that reactive maintenance -- "fix it when it breaks," is costly, inefficient and in many cases unsafe. That led to preventive maintenance, repair or replace at regular scheduled intervals less than the average time to failure.

In many cases preventive maintenance also missed the mark. Much preventive maintenance proved unnecessary. Components were replaced on an arbitrary schedule, often in excellent condition with substantial lifetime remaining. Horror stories abound where performance and condition degraded, and major problems occurred, following unnecessary preventive maintenance. Enter predictive or condition directed maintenance. Maintenance tasks are based on condition measurements and performed on the basis of defects -- before outright failure impacts safety and production. Predictive maintenance has proven very successful -- with two possible exceptions. Ongoing expenses are relatively high. Few have successfully translated technical and operating results to value and benefits in the financial terms necessary to assure continued support. As a result, in industry many successful predictive maintenance programs are being curtailed, even terminated, as cost reduction measures.

Cost reduction has several added dimensions. One is the loss of experience as senior maintenance workers and supervisors take advantages of incentives for early retirement. Today, cost reduction is primarily a numbers game. Only a few recognize the connection between safe, reliable operation and the absence of problems with the contribution, commitment and experience of individuals responsible for the results. With the departure of champions and their programs being dismantled, what will happen? Some say a time bomb is ticking on the momentum of past efforts. Statistics of increasing reportable incidents across industry have been cited as first evidence of a serious problem.

GROUND HOG DAY -- BUT IS SPRING JUST AROUND THE CORNER? There is a growing realization that the entire idea of maintenance must be rethought and reconstructed to fit within a twenty-first century environment. Characteristics include more intense operation, greater consequences of failure, fewer people and less experience. Within this environment, maintenance must become part of a larger process to assure safety and availability.

The ability to meet mission requirements at full capacity, efficiency, safety and quality is not just an objective but the objective. Benefits and value must be expressed in terms that connect directly to organizational business and operating objectives. They must be relevant, credible and compelling to senior executives. For the purposes of this paper the concept will be called Asset Management. The name is unimportant. The concept is vital.

BEYOND MAINTENANCE TO ASSET MANAGEMENT: Asset management is an overall concept that goes well beyond maintenance. Asset management includes all the critical factors that establish and influence the ability to accomplish a specific mission safely and at full capacity. Critical factors include design, specification, installation, safety, economics, operation, risk, surveillance, information, maintainability, work practices and training. Each fulfills a vital role in the overall concept.

Asset management extends the craftsman credo, "perform a task right the first time," to completely eliminate the necessity for task whenever possible. And, this is something everyone must understand. ***The only real way to gain a permanent reduction in maintenance is to reduce the need for maintenance.*** The philosophy begins at design and continues throughout equipment lifetime. Design for the service, construct using suitable materials, require proper installation, pay attention to lubrication and lubrication integrity, enforce proper repair procedures and eliminate chronic problems. The list is familiar to experienced professionals. Demonstrated results are as close as your automobile showroom. Automobile maintenance requirements have been dramatically reduced over the past ten years -- and reliability dramatically increased. How? Problems have been designed out.

Successful implementation of this philosophy depends on two factors; financial justification and immediate access to information.

ECONOMICS -- LOVE IT BUT NEVER IGNORE IT: In the financial area, improved practices must be supported with compelling financial justification. Financial justification must be in terms that are familiar and credible to senior executives. Here again the craftsman credo of performing a task right is probably justified financially -- when all the benefits and effects are fully accounted.

What are the best standards for measuring maintenance performance? Is it reliability? How does reliability relate to performance of a given mission, especially if required availability is significantly less than 100% and/or there are system redundancies? Some suggest a measure of the proportion of scheduled to total maintenance. But what value is created by a high percentage of scheduled to total maintenance if the scheduled tasks are unnecessary or improperly performed? (OEE) Overall Equipment Effectiveness has been proposed[1,2]. How about cost per unit output ($/MW)? RONA (Return On Net Assets) and NPV (Net Present Value) are under discussion. Whatever the selection for maintenance benchmarks they must connect directly to unit objectives and economics. Nothing else will attract the attention of the executives who control the funds!

INFORMATION -- THE ENGINE: If financial justification is the wing that lifts the airplane, information is the engine that drives it forward. And not just any information.

Current and projected performance and condition, equipment specifications, operating and repair history, operating and maintenance recommendations and task instructions, including safety precautions are all necessary. Imagine if you will Internet like immediate access to comprehensive information about any given piece of equipment, all from a single display. What is it doing now? What is it supposed to do? Have there been problems in the past? How were they resolved? How to get from here to there? Combine all this with detailed operating and maintenance instructions and instant communications everywhere in the world. Fewer, highly experienced people can accomplish more with greater efficiency. In the financial world this is called leverage.

And how is maximum power and efficiency obtained from the information engine? A growing number of people are concluding that the answer is self integrating, open information systems. The same philosophy that has driven the success of the personal computer far beyond what anyone could have imagined only ten short years ago. The same concept that provides total access to an incomprehensible array of information on the Internet. Pay an inexpensive connection fee and the gate into the information superhighway opens wide.

What will this mean for Asset Management? Free information exchange of mechanical condition, performance, operating and maintenance records between condition assessment, maintenance management and process control systems. A maintenance mechanic can see a machine's condition, how it is operating compared to how it ought to be operating and any changes that have occurred. There will be full access to history and task instructions, including parts lists, and safety precautions. If a problem occurs, an analyst can recover measurements and the exact sequence of events leading up to and during the problem. If that sounds too far into the future, most modern automobiles provide this same capability in their on-board computers. Mechanical problems are immediately and accurately diagnosed by unloading the computer.

MIMOSA -- THE PATH TO OPEN ASSET INFORMATION SYSTEMS: All available evidence supports the superiority of open systems. In the world of PC's the current, highly publicized, travails of Apple Computer have their origin, at least in part, to the decision to remain closed. Despite arguably inferior technology compared to Apple, during the early, introductory years, the open IBM architecture gained overwhelming commercial superiority. Process control is rapidly moving from closed proprietary systems to open access and exchange. Why? A growing customer awareness that no single supplier can maintain "best for application" in every area required for modern automated control systems.

MIMOSA, The Machinery Information Management Open Systems Alliance, is rapidly becoming a rallying point for open asset management systems. MIMOSA has been meeting for approximately 2 1/2 years. CRIS (Common Relational Information Schema), the MIMOSA convention for exchange basic vibration information, is complete and was tested successfully in December 1996 by a multi supplier demonstration conducted during the National Predictive Maintenance Technology Conference. The open convention is published on the Internet, along with the MIMOSA charter and current information, at **www.hsb.com/pcm/mimosa**.

MIMOSA primary objectives:

- Formulate and publish *consensus* conventions for information exchange between:
  - condition assessment systems; vibration, lube oil, performance, thermography, electrical characteristics
  - condition assessment and maintenance management (CMMS), control (DCS) and other asset information and management systems
- Provide a neutral forum for vital dialog between progressive information system purchasers, suppliers and integrators

Currently, the MIMOSA information exchange convention is being enlarged to include equipment information needed throughout an organization. This includes current status, problem identification and severity, time to action, components affected, operating and maintenance recommendations[3,4].

When complete, MIMOSA will provide full, open access to all vital asset information throughout an organization without the necessity to construct expensive, inflexible and difficult to maintain proprietary interfaces between specialized information sources and users. Purchasers gain the benefit and value of components that are best for their specific application, mutually reinforcing and capable of full integration with enterprise information without self limiting, inflexible and costly custom links.

Sound like utopia? Perhaps, but others have shown that it can be done. The journey has begun for equipment asset management information.

CONCLUSION: The entire concept of maintenance is at a crossroads. Continue as headed, concentrate on keeping equipment operating and likely revert to maintain on failure. The alternative -- go beyond maintenance to asset management, seek value, provide financial justification and become an indispensable factor in effectively achieving the organization's mission. Which will it be? The first requires no effort. The second requires a vision, strategy, initial and continuing proof-of-concept and a lot of hard work convincing seniors that the non obvious is the best. The choice is up to you!

REFERENCES:

1. Robinson, Charlie, "Calculating Process Overall Equipment Effectiveness," *Maintenance Technology,* June 1994.
2. Robinson, Charlie, "Calculating Plant Overall Equipment Effectiveness," *Maintenance Technology,* November 1994.
3. Mitchell, John, "MIMOSA -- One Year Later," *Sound & Vibration,* September 1996.
4. Baldwin, Robert C., "Model for Maintenance Information Exchange," *Maintenance Technology,* December 1996.

MFPT

FAILURE ANALYSIS

CoChairs: **Victor K. Champagne**
U. S. Army Research Laboratory

Marc Pepi
U. S. Army Research Laboratory

INVESTIGATION OF MAIN ROTOR SPINDLE ASSEMBLY RETAINING RODS FOR AN ARMY UTILITY HELICOPTER FABRICATED BY THREE DIFFERENT CONTRACTORS

Marc Pepi
Victor Champagne
Scott Grendahl
US Army Research Laboratory
Weapons and Materials Research Directorate
AMSRL-WM-MD
Aberdeen Proving Ground, MD 21005-5059

Dr. Kirit Bhansali
US Army Aviation and Troop Command
AMSAT-R-EFM
St. Louis, MO 63120-1798

Abstract: The US Army Research Laboratory (ARL) was tasked by the US Army Aviation and Troop Command (ATCOM) to perform a metallurgical examination of main rotor spindle assembly retaining rods fabricated from precipitation hardened (PH) 13-8 Mo stainless steel by three different manufacturers ("A", "B" and "C"). These components were subjected to prior spectrum load fatigue testing in order to qualify an alternate source. Manufacturer "A" was the current source, and this testing was performed to compare the results of "B" and "C" versus "A". The "C" components exhibited only half the fatigue resistance of the "A" and "B" components. Upon dimensional examination by ARL, it was determined that the root radius of an "A" and "C" rod did not conform to the governing requirements. Although the surface roughness of the parts from each manufacturer varied, all readings met the governing requirement. Hardness results differed only slightly between the rods, however, the tensile properties (of coupons sectioned from the original rods) were significantly different. Nevertheless, the hardness and tensile results met the minimum requirements specified. The results of fatigue testing (of coupons sectioned from the original rods) showed a dramatic difference between the rods. Metallography was utilized to examine the microstructure and grain size. The structure of each rod was consistent with the prior treatment, and the grain size met the governing requirement. The amount of delta (free) ferrite within the structure varied slightly from rod to rod, but was well within the specified limits. Banding was more prevalent within the "C" material than the "A" and "B" material. The threads of each rod were examined metallographically, since this was the area of failure as a result of the spectrum load fatigue testing. Although differences in the surface profile of the threads from the different manufacturers was noted, there was no evidence of gross abnormalities such as tear out or chatter. The chemical analysis of each rod varied, but each composition met the governing requirements. ARL subsequently heat treated a

section from the "A" rod and a section of the "C" (the material from these two manufacturers showed the greatest disparity in ARL fatigue results) and machined mechanical test specimens (tensile and fatigue). This heat treatment included a solution heat treatment at 1700°F for 1 hour plus an aging cycle performed at 1050°F for 4 hours. These specimens yielded very similar hardness, mechanical property and fatigue resistance results. The hardness and mechanical properties of the "A" material actually had been slightly reduced, while the same attributes of the "C" material had risen considerably. It was determined that even though the hardness and tensile properties of the "C" material had met the minimum requirements of the governing specification, the values were not "typical" of PH 13-8 Mo steel aged at 1050°F. Based upon the results of re-heat treating, it was concluded that an inadequate prior heat treatment sequence was performed by the manufacturer.

Key Words: Failure analysis; heat treatment, mechanical testing; PH 13-8 Mo stainless steel.

Spectrum Load Fatigue Testing: A total of six retaining rod components were spectrum load fatigue tested at the US Army Aviation Applied Technology Directorate (AATD), Ft. Eustis, VA. A typical retaining rod is shown in Figure 1 in the as-received condition. Manufacturer "A" was the original contractor. Retaining rods from manufacturers "B" and "C" were tested and compared to the results of Manufacturer "A" as part of qualification testing. The results of this component testing are listed in Table I. The investigation performed by ARL was necessitated due to the disparity in these fatigue results between "A" (and "B") and "C"

Table I. Spectrum Load Fatigue Results
Testing at AATD

Specimen	# of Cycles
A-1	108,460
A-2	157,274
B-1	94,335
B-2	131,460
C-1	46,735
C-2	54,637

Governing Specifications:

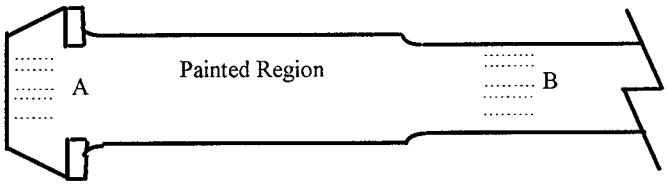
AMS 5629, "Steel, Corrosion Resistant, Bars, Forgings, Rings, and Extrusions."
AMS 2315, "Determination of Free Ferrite Content."
MIL-S-8879, "Screw Threads, Controlled Radius Root With Increased Minor Diameter."

Dimensional Inspection: The threads of each component ("A", "B" and "C") were inspected utilizing an optical comparator. This region was scrutinized since this was the location where each rod failed during AATD component testing. According to the governing specification, the thread dimensions should conform to MIL-S-8879. *Table III, Fine Thread Series* of MIL-S-8879 references the dimensional acceptance criteria for thread characteristics. Table II outlines the acceptance criteria for the threads along with the data acquired from rods for the external thread - UNJF Class 3A. The values indicated are the average of three measurements obtained. Rods A-1 and B-1 were out of tolerance on the root radius.

Table II. Dimensional Analysis Data

Specimen Number	Basic Size	Threads Per Inch	Major Dia. (in)	Pitch Dia. (in)	Minor Dia. (in)	Root Rad. (in)
A-1	1.25	12	1.2455	1.1953	1.1463	0.0108
A-2	1.25	12	1.2464	1.1943	1.1487	0.0135
B-1	1.25	12	1.2429	1.1951	1.1511	0.0146
B-2	1.25	12	1.2409	1.1941	1.1501	0.0146
C-1	1.25	12	1.2491	1.1945	1.1480	0.0217
C-2	1.25	12	1.2487	1.1951	1.1491	0.0142
MIL-S-8879	1.25	12	1.2386 min. 1.2500 max.	1.1913 min. 1.1959 max.	1.1442 min. 1.1538 max.	0.0125 min. 0.0150 max.

Surface Profile: The surface profile of each of the retaining rods was measured utilizing a Mitutoyo Surftest Analyzer in two areas along the part in the locations noted below:



The following table presents the data obtained during surface profile testing. The data was measured as RMS, in μinches. Three readings were taken at each location for each rod.

The engineering drawing requires a surface finish of 125 μinches maximum. The results in Table III show that although the "B" rods had a coarser surface finish than the "A" and "C" components, the requirements were met for each rod.

Table III. Surface Finish Results (RMS μ inches)

Specimen	A	Average	B	Average
A-1	25. 26. 25	25.3	25. 21. 18	21.3
A-2	15. 13. 21	16.3	8. 12. 16	12.0
B-1	60. 60. 66	62.0	52. 45. 54	50.3
B-2	100. 104. 104	102.7	87. 91. 83	87.0
C-1	43. 43. 43	43.0	21. 30. 34	28.3
C-2	45. 46. 47	46.0	10. 14. 16	13.3
		Rqmt. <125		Rqmt. <125

Mechanical Testing - Hardness: Specification AMS 5629 required a minimum hardness of 40 HRC for the H1050 heat treat condition for the PH 13-8 Mo material. Surface macrohardness was determined from Rockwell "C" scale measurements performed directly on the part utilizing a 150 kg major load. A total of ten readings were obtained from each specimen. The data acquired was corrected according to the standard round work correction for cylindrical specimens. This data is presented in Table IV. The hardness values conformed to the governing specifications. The average hardness value of the "B" components fell between the values of the "A" and "C" samples. However, the "B" hardness values more closely resembled the "A" values rather than the "C" values.

Table IV. Macrohardness (Corrected)
HRC Hardness Values 150 Kg Major Load

A-1	A-2	B-1	B-2	C-1	C-2
41.9	42.2	41.7	42.5	40.7	40.4
41.9	42.3	41.0	42.7	40.9	40.9
43.4	40.7	41.7	41.2	40.6	41.3
41.2	42.3	42.0	40.8	40.3	41.1
42.5	41.2	39.8	42.5	40.8	41.1
41.8	40.6	41.7	42.2	40.2	40.7
40.8	42.3	40.3	41.1	40.4	39.8
42.4	42.0	41.1	41.4	40.5	40.7
43.0	42.2	41.9	42.2	40.4	40.0
43.4	42.1	41.6	43.0	40.2	40.5
Avg. = 42.2 HRC	Avg. = 41.8 HRC	Avg. = 41.2 HRC	Avg. = 42.0 HRC	Avg. = 40.5 HRC	Avg. = 40.7

AMS 5629 H1000	43 HRC minimum
AMS 5629 H1025	41 HRC minimum
AMS 5629 H1050	40 HRC minimum
AMS 5629 H1100	34 HRC minimum
Typical H1000 [1]	45 HRC
Typical H1050 [1]	40 HRC
Typical H1100 [1]	35 HRC

Mechanical Testing - Tensile: Specification AMS 5629 outlines the requirements for the tensile properties of PH 13-8 Mo material aged to the H1050 condition. Two tensile specimens were fabricated from each of the retaining rods in order to confirm that the prior H1050 heat treatment was performed properly. Cylindrical rods were sectioned from the components by wire EDM, and final machining was performed to achieve the dimensions of a standard ASTM round, threaded tensile specimen. Table V contains the results of ARL-MD tensile testing. The minimal values of this alloy in the H1000, H1025 and H1050 condition as listed within AMS 5629 are also provided. In addition, the typical tensile properties of the H1000 and H1050 conditions as received from Armco Steel Corporation [1] are listed. The results indicate that the "A" and "B" material had been heat treated to the H1050 condition meeting both the minimal tensile properties of AMS 5629 and the typical properties established by Armco steel. The "C" specimens met the minimal values of the tensile properties of H1050, but the values were much lower than the typical values of this condition. This difference in tensile properties may be related to the difference in fatigue properties noted both from the actual components and the specimens fabricated by ARL-MD (see Fatigue section).

Table V. ARL Tensile Testing Results

Specimen	Area (sq. in.)	0.2% YS (psi)	UTS (psi)	% RA	% EL
A-1	0.0499	201.000	203.600	64.6	14.0
A-2	0.0495	180.000	189.200	69.8	15.0
B-1	0.0495	185.100	195.800	65.7	13.8
B-2	0.0495	185.300	195.900	66.2	13.2
C-1	0.0491	178.700	180.300	65.0	16.0
C-2	0.0495	178.200	182.700	67.6	14.4
AMS 5629 H1000	(Long.)	190.000	205.000	50.0	10.0
AMS 5629 H1025	(Long.)	175.000	185.000	50.0	11.0
AMS 5629 H1050	(Long.)	165.000	175.000	50.0	12.0
Typical H1000 [1]	(Long.)	205.000	215.000	55.0	13.0
Typical H1050 [1]	(Long.)	180.000	190.000	55.0	15.0

Mechanical Testing - Fatigue: Fatigue specimens were fabricated by the same machinist utilizing the same process and tooling in order to minimize the effects of machining. This concern was a result of the initial fatigue tests performed on actual components in which the failures occurred within the threads of each component. This raised questions concerning the thread dimensions, surface finish, etc., since three different manufacturers were involved in producing the parts. One fatigue specimen was fabricated from each retaining rod, to the dimensions shown in Figure 2. Testing was conducted on a Model 8502 Instron servohydraulic test machine, with an R value of 0.1 and a frequency of 25 Hz. As dictated by ATCOM, a stress level of 170 ksi was utilized. The data points obtained showed that the "A" and "B" material did not fail after a total of 3 million cycles. The "C" specimens, however, failed at 83,020 and 168,567 cycles, respectively.

Table VI. ARL Fatigue Testing Results

Specimen	Area (sq. in.)	Stress (ksi)	R	Frequency (Hertz)	Cycles to Failure
A-1	0.0184	170	0.1	25	3,000,000+
A-2	0.0184	170	0.1	25	3,000,000+
B-1	0.0184	170	0.1	25	3,000,000+
B-2	0.0184	170	0.1	25	3,000,000+
C-1	0.0184	170	0.1	25	168,567
C-2	0.0184	170	0.1	25	83,020

Metallography - Structure, Grain Size: Alloy PH 13-8 Mo is a martensitic precipitation hardenable stainless steel with a composition balanced to prevent the formation of delta ferrite in the structure It has low carbon content to minimize grain boundary carbide precipitation and is double vacuum melted to reduce alloy segregation [2]. AMS 5629 lists two requirements with respect to the microstructure of this alloy:

3.4.1.2: Microstructure: The product shall contain not more than 2% free ferrite, determined in accordance with AMS 2315,

3.4.2.1.3: Grain Size: Shall be 5 or finer for product up to 3.00 inches in nominal cross sectional thickness with occasional grains as large as 3 determined by comparison of a polished and etched specimen with the chart in ASTM E 112.

Metallographic samples representing each component were mounted in phenolic powder with edge retention, and rough polished with silicon carbide papers ranging in grit from 240 to 600. The samples were subsequently fine polished utilizing 3 and 1 micron diamond suspension, followed by 0.05 micron colloidal silica. The structure of the samples was consistent with vacuum induction plus consumable electrode melted material in that there was no evidence of gross inclusions or anomalies in the as-polished condition. The samples were subsequently etched using Vilella's reagent, to compare the microstructural features. At low magnification (50x), varying degrees of banding was noted within the structure of the each samples. The banding noted within the "A" and "B" material was slight, compared to the heavy banding noted within the "C" material. Figure 2 shows an example of the banding in the "C" material. Note the free ferrite within the bands (indicated by arrows). At higher magnification (100x), stringers of free (delta) ferrite was observed within the bands of each material. An example of this ferrite is shown in Figure 3. The delta ferrite is not permitted to exceed 2% in this alloy, since the interface between these and the matrix provide sites for segregation and crack propagation. Material "B" exhibited the least amount of free ferrite, followed by the "A" and "C" . The "C" material exhibited the heaviest amount of free ferrite. It should be noted that although the amount of free ferrite varied, each sample was well below the maximum allowable. The AMS 5629 requirement of 2% maximum free ferrite in the material was easily met by the retaining rods from each manufacturer.

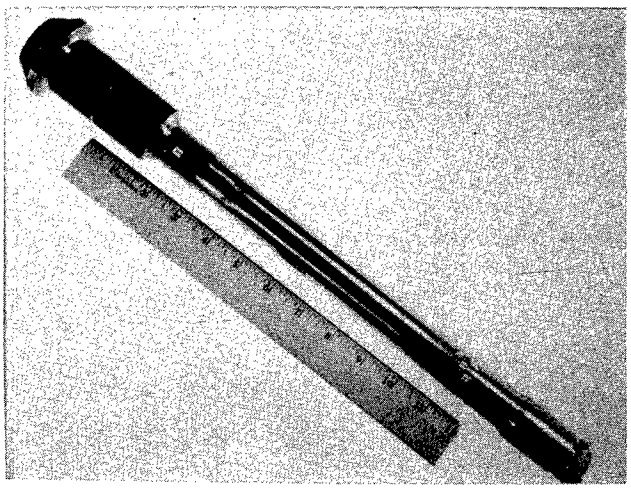


Figure 1 Typical retaining rod subject to spectrum load fatigue testing in the as-received condition. Reduced 75%.

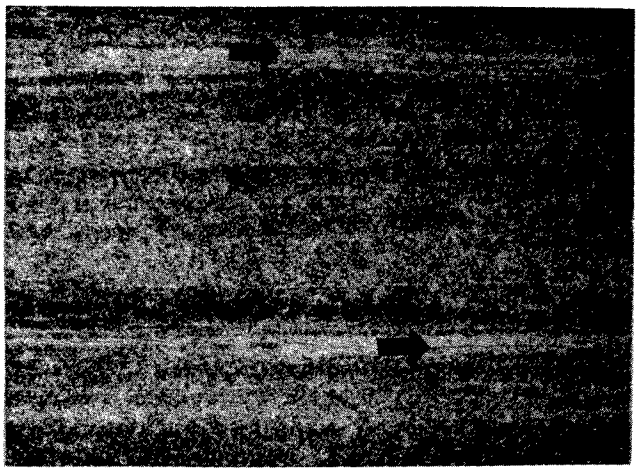


Figure 2 Banding noted within the structure of the manufacturer "C" material. Note the delta ferrite within the lightly shaded bands (indicated by arrows). Banding was less prevalent in the "A" and "B" material. Villela's reagent. Mag. 50x.

At higher magnification (1000x), the structures of each sample was compared in both the longitudinal and transverse directions. An attempt was made to characterize the heat treatment by microstructural analysis. However, as confirmed earlier by Armco Steel, there is not much microstructural difference between the H1025, H1050 and H1100 conditions [3]. The structure of each material was similar, consisting of precipitated compounds within a martensitic matrix. The grain size was also similar in size and geometry in both the longitudinal and transverse directions (Figure 4). The grain size of each sample was much finer than the "5" required by AMS 5629. The uniformity of the machined surface of the threads was also examined metallographically. The machined surface of the "B" threads were by far the smoothest while the that of "C" were the roughest. Examination of the threads using scanning electron microscopy revealed no evidence of chatter or gross abnormalities such as tear out on any of the rods.

Chemical Analysis: Chemical Analysis was performed to verify the elemental composition of the each retaining rods. The chemical composition was governed by AMS 5629. The designated material was PH 13-8 Mo, heat treated to the H1050 condition. Inductively coupled plasma/atomic emission spectroscopy was utilized to determine the weight percentage of each element. Samples were sectioned from each rod utilizing a diamond specimen saw. Table VII lists the results of this testing. Each sample conformed to specification AMS 5629. Unlike the "A" and "C" rods, there were no trace elements found in the "B" rods. Although the element compositions were all within specification, the "B" rods had slightly higher amounts of manganese, silicon, chromium, nickel and molybdenum. The overall effect of a higher composition of these particular elements is an increase in the tensile strength and hardness of the material. It is indeterminate whether these small increases would have a significant impact on the overall performance of the components under fatigue, but the "B" rods did have a higher hardness and tensile strength than the "C" rods.

Table VII. Chemical Composition

Element	A-1	A-2	B-1	B-2	C-1	C-2	AMS 5629
Carbon	0.050	0.034	0.040	0.040	0.040	0.043	0.05 max.
Manganese	<0.01	<0.01	0.05	0.04	0.03	0.03	0.10 max.
Silicon	<0.01	<0.01	0.07	0.06	<0.01	<0.01	0.10 max.
Chromium	12.38	12.32	12.65	12.71	12.36	12.47	12.25-13.25
Nickel	7.93	7.82	8.21	8.27	7.94	7.81	7.5-8.5
Phosphorous	0.007	0.008	0.008	0.008	0.010	0.010	0.01 max.
Sulfur	0.002	0.002	<0.005	<0.005	0.001	0.001	0.008 max.
Molybdenum	2.15	2.09	2.18	2.17	2.11	2.06	2.0-2.5
Aluminum	1.05	1.08	0.96	1.02	1.14	1.15	0.9-1.35
Nitrogen	0.003	0.004	0.003	0.004	0.005	0.004	0.01 max.
Copper	<0.01	0.02	<0.01	<0.01	0.01	0.01	-
Vanadium	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-
Titanium	0.01	0.02	<0.01	<0.01	0.01	0.01	-
Niobium	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-
Cobalt	<0.01	<0.01	<0.01	<0.01	0.06	0.06	-
Magnesium	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-
Iron	rem.	rem.	rem.	rem.	rem.	rem.	rem.

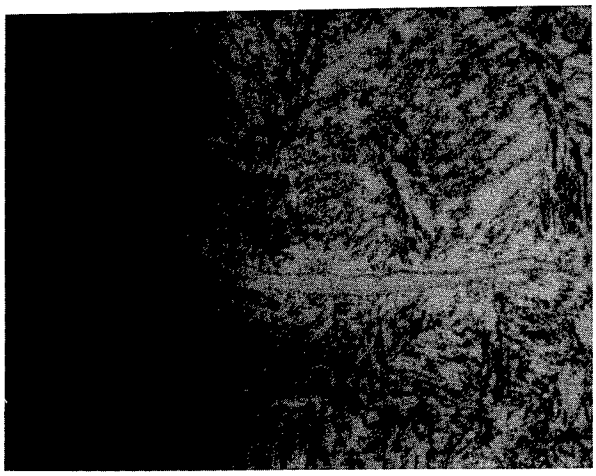


Figure 3 Enlargement of a delta ferrite stringer noted within the "C" material. Delta ferrite stringers were noted in the "A" and "B" material to a lesser extent. Villela's reagent. Mag. 1000x.

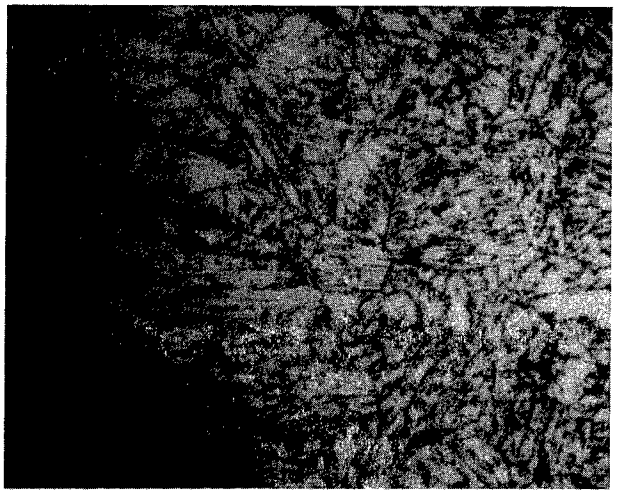


Figure 4 Typical grain size noted within the structure of the "C" material. This grain size was similar to that of the "A" and "B" samples. Villela's reagent. Mag. 1000x.

Heat Treatment: Sections of an "A" and "C" rod were re-heat treated and aged to the H1050 condition at ARL, with the intent of determining whether the prior heat treatment was performed adequately. The following procedure was utilized (from AMS 5629):

Solution Heat Treat:

*Heat to 1700°F +/- 25°F for one hour minimum,
*Air cool until the parts are warm enough to handle,
*Run samples under cold water.

Age:

*Heat to 1050°F +/- 10°F for 4 hours +/- 15 minutes,
*Air cool.

Specification AMS 5629 required a minimum hardness of 40 HRC for the H1050 heat treat condition. Surface macrohardness was determined from Rockwell "C" (HRC) scale measurements taken directly on the heat treated parts. This scale utilized a 150 kgf major load. A total of ten readings were obtained from each of the two rods. The data acquired was corrected according to the standard round work correction for cylindrical specimens. The data in Table VIII shows a slight decrease in the hardness of the "A" material, and a significant increase in the hardness of the "C" material

Table VIII. Hardness After ARL Heat Treatment
Rockwell "C" Scale, 150 kgf major load

A	C
42.3	42.5
42.7	43.4
42.1	43.2
42.9	43.5
42.6	43.7
43.2	43.7
42.9	43.9
42.8	42.7
42.2	43.6
43.1	43.5
Avg. = 42.7	Avg. = 43.4
Avg. of "A" Rod Before HT = 42.0	Avg. of "C" Rod Before HT = 40.6

A tensile specimen and two fatigue specimens were fabricated from each re-heat treated rod to confirm the mechanical properties of the H1050 condition. The 0.2% yield strength and the UTS of the "A" material remained consistent as a result of the heat treatment, while these properties were significantly greater for the "C" material. The properties of the re-heat treated "A" material closely resembled those of the re-heat treated "C" material. The tensile results are shown in Table IX.

Table IX. Tensile Properties Before and After ARL Heat Treatment

Specimen	0.2% Y.S. (psi)	UTS (psi)	%RA	%El
A - Before	201,000	203,600	64.6	14.0
A - Before	180,000	189,200	69.8	15.0
A - After	190,200	198,000	69.0	13.0
C - Before	178,700	180,300	65.0	16.0
C - Before	178,200	182,700	67.6	14.4
C - After	189,900	197,500	66.3	15.0

Fatigue testing was performed on a Model 8502 Instron servohydraulic test machine with an R value of 0.1 and a frequency of 25 Hz. As dictated by ATCOM, a stress level of 170 ksi was utilized. Table X lists the fatigue test data of the re-heat treated specimens, and the previous results obtained from the material in the as-received condition. The results show that after re-heat treatment, the "C" material ran out to 3 million cycles, as did the "A" material.

Table X. Fatigue Properties Before and After ARL Heat Treatment

Specimen	Stress (ksi)	R Value	Frequency (Hertz)	Cycles to Failure
A - Before	170	0.1	25	3+ million
A - Before	170	0.1	25	3+ million
A - After	170	0.1	25	3+ million
A - After	170	0.1	25	3+ million
C - Before	170	0.1	25	168,567
C - Before	170	0.1	25	83,020
C - After	170	0.1	25	3+ million
C - After	170	0.1	25	3+ million

Discussion - ARL Tension and Fatigue Testing: There was a general correlation between the tensile properties and fatigue resistance between the three manufacturers. Those specimens which had a higher tensile strength and hardness ("A" and "B") also showed greater values of fatigue in both component and laboratory specimen fatigue testing. The material from each manufacturer satisfied both the minimal and typical values for tensile properties as listed in Table V. The tensile and hardness data indicated that the "C" material, although satisfying the minimal requirements for yield strength and UTS falls short of satisfying the "typical" values for the H1050 condition, listed within Table V. These "typical" values were provided to ARL-MD by Armco Steel Corp., the developer of the PH 13-8 Mo alloy. The "A" material not only meets the minimal and typical values of UTS, but of the two tensile specimens tested, one actually exhibited a UTS value 13 ksi higher than the typical value and fell close to the minimal requirement for the H1000 condition.

Additional tension and fatigue testing would be required to obtain a greater understanding of the exact relationship between the difference in UTS and fatigue behavior observed in each material. However, it can be stated from the data obtained, that material which only satisfied minimal tensile properties were most likely to fail the spectrum load fatigue criteria, while material meeting the typical tensile properties were most likely to pass this criteria.

Discussion - Microstructural Differences: The "A" and "B" rods contained limited banding and delta ferrite stringers, while the "C" material contained a greater amount of each. The material specification AMS 5629 limits the amount of delta ferrite to 2%. Ferrite can adversely affect the fatigue resistance of this alloy in concentrations greater than 5%. However, it has been deduced by metallography that the amount of free ferrite in the each rod representing the three manufacturers was much less than 2%. The surface finish differed between each manufacturer in the threaded region. These differences may also account for differences in fatigue behavior. The magnitude of its effects are not quantifiable by metallurgical evaluation.

Discussion - ARL Heat Treatment: A section of an "A" and "C" retaining rod were heat treated and aged at ARL. The purpose of this task was to determine whether the re-heat treatment of these rods would affect the mechanical properties, primarily fatigue. It was anticipated prior to this heat treatment that the "A" material would be unaffected, while the "C" material would experience an increase in tensile and fatigue properties. In the as-received condition, the "A" material exhibited properties exceeding the typical properties of the H1050 condition, while the "C" material met the minimum, but not typical requirements of the H1050 condition. In fact, the "A" material was close to the minimal H1000 properties. Subsequent testing on each re-heat treated rod showed an increase in hardness, tensile strength and fatigue resistance for the "C" material, and a corresponding slight decrease in these characteristics for the "A" material. Based upon this information, it was determined that heat treatment played a significant role in the premature spectrum load fatigue testing failures of the "C" material.

Conclusions: Although the material from each manufacturer met the minimum mechanical property requirements of PH 13-8 Mo in the H1050 condition, only the material from "A" and "B" met the typical requirements of this alloy. There were no significant differences in dimensional tolerance, surface finish, microstructure, grain size or chemical composition. As a result of re-heat treatment by ARL, the properties of the "C" material increased dramatically. Therefore, it was concluded that the "C" material was not adequately heat treated which was attributable to low spectrum load fatigue resistance.

Future Work: In order to obtain a greater understanding of the effects of prior heat treatment, a number of "C" retaining rods will be re-heat treated in accordance with the governing specification, and subjected to spectrum load fatigue testing.

References

[1] Armco PH 13-8 Mo Stainless Steel (XM-13, UNS S13800), Product Data Bulletin No. S-24.

[2] "Mechanical Property Data, PH 13-8 Mo Stainless Steel", Battelle Memorial Institute, Columbus, OH, 1972, p. 41.

[3] Conversation with Mr. Joseph Jasper, Armco Corporation, 15 May 1996.

DYNAMICS OF LAYERED COMPOSITE BEAMS AND PLATES WITH A SINGLE DELAMINATION

Marek Krawczuk, Wiesław Ostachowicz, Arkadiusz Żak
Institute of Fluid Flow Machinery, Polish Academy of Sciences
Fiszera 14, 80-952 Gdańsk, Poland

Abstract: In this work a model of a layered composite beam and plate with a single delamination failure is presented. Undamaged parts of the beam and plate are modeled by finite elements while the delamination is substituted by special boundary conditions. The elaborated model can be applied for the analysis of the influence of the delamination length and location on changes in bending natural frequencies and mode shapes of laminated composite beams and plates. Information about dynamic behavior of delaminated composite beams and plates received from the elaborated models can be used in diagnostic systems based on vibration measurements.

Key Words: Delamination; layered composite; natural frequencies; mode shapes.

INTRODUCTION: The use of composite anisotropic materials in the fields of mechanical and civil engineering has substantially increased over the past few years. These materials are particularly widely used in situations where a large strength-to-weight ratios are required. Composite materials similarly to isotropic materials are subjected to various damages, mostly cracks and delaminations. Delamination is one of the most important failure modes of laminated composite materials. They result in local changes of the stiffness of the element and consequently its dynamic characteristics are altered. Changes of natural frequencies and mode shapes, amplitudes of forced vibrations and also coupling of vibrations forms are observed. The dynamic characteristics of damaged elements can be correlated with the location and size of damages. These relations are frequently used in diagnosis of such constructional elements – for example [1–2].

This work presents the method of modeling delaminations in the composite beams and plates based on the finite element formulation [3–5]. Elaborated models enable the analysis of the influence of the delamination (i.e. its length and location) and also material properties on the dynamic characteristics of constructions made of composite materials. In this work the beam is modeled by beam finite elements with three nodes and three degrees of freedom at each node (i.e., two axial displacements and one independent rotation) while the plate is modeled by plate finite elements with eight nodes and five degrees of freedom at each node (i.e., three axial displacements and two independent rotations). In the delaminated region additional boundary conditions are applied. It is assumed that the delamination extends to the full width of the element and remains open (i.e. contact forces between lower and upper parts are neglected in the model). Also an easy modification of

these elements is possible, according to the analysis of the specific damage (i.e. multiple delaminations). The elaborated models are used in the investigations of the influence of the delamination length and position, material properties, number of material layers on changes in bending natural frequencies.

DELAMINATED BEAM FINITE ELEMENT: A discrete model of a delaminated part of the beam is presented in the Fig.1. The delaminated region is modeled by three beam finite elements (I, II and III) which are connected at the tip of the delamination by additional boundary conditions.

The layers are located symmetrically with respect to the x–z plane. Each element has three nodes at x=-L/2, x=0, x=L/2. At each node there are three degrees of freedom which are axial displacements q_1, q_4, q_7, transverse displacements q_3, q_6, q_9 and the independent rotations q_2, q_5, q_8. Additionally, it is assumed that the number of degrees of freedom is independent of the number of layers.

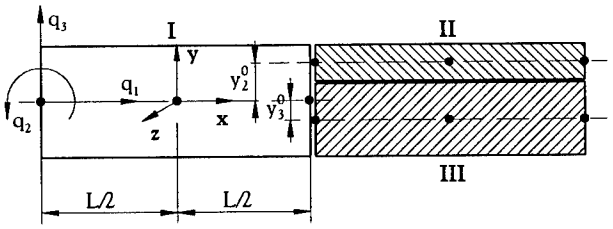


Fig.1. The delaminated region of a beam modeled by finite elements.

Neglecting warping, the displacements u and v of a point can be expressed as:

$$\begin{cases} u(x,y) = u^0(x) - y \cdot \phi(x) \\ v(x,y) = v^0(x) \end{cases} \tag{1}$$

where $u^0(x)$ denotes the axial displacement, $\phi(x)$ the independent rotation, and $v^0(x)$ the transverse displacement.

In the finite–element modeling, the bending displacement $v^0(x)$ is assumed to be cubic polynomial in x, while the axial displacement $u^0(x)$ and the rotation $\phi(x)$ are assumed to be quadratic. Additionally, it is assumed that the shear strain variation is linear, as proposed by Tessler and Dong [6]. Employing the above conditions, the displacements and rotation in the element may be written in the following form:

$$\begin{cases} u^0(x) = a_1 + a_2 x + a_3 x^2 \\ \phi(x) = a_4 + a_5 x + 3a_9 x^2 \\ v^0(x) = a_6 + a_7 x + a_8 x^2 + a_9 x^3 \end{cases} \tag{2}$$

The coefficients a_1–a_9 can be expressed in terms of the element degrees of freedom by using the nodal conditions and next the matrix of the shape functions for the single layer can be determined.

Employing the shape functions matrix for the single layer, the inertia matrix of the whole element may by calculated as a sum of all single layer inertia matrices. In the same way the stiffness matrix for the whole element can be obtained when the strain–displacement matrix for each layer is evaluated.

Boundary conditions at the tip of the delamination: In order to connect elements modeling the delaminated area of the beam (I, II and III) the following boundary conditions are applied at the tip of the delamination:

$$\begin{cases} u_1^0(x) = u_2^0(x) + y_2^0\phi_2(x) & \phi_1(x) = \phi_2(x) = \phi_3(x) \\ u_1^0(x) = u_3^0(x) + y_3^0\phi_3(x) & v_1^0(x) = v_2^0(x) = v_3^0(x) \end{cases} \tag{3}$$

where y_2^0 and y_3^0 denote distances between neutral axes of elements I–II and I–III, respectively (see Fig. 1).

Using the relations for coefficients a_1–a_9 for a single element layer and taking into account relations (3), it can be found that coefficients a_1–a_9 for elements I, II and III must satisfy the following relations:

$$\begin{cases} a_1^{II} = a_1^I - y_2^0 a_4^I & a_1^{III} = a_1^I - y_3^0 a_4^I & a_4^I = a_4^{II} = a_4^{III} & a_7^I = a_7^{II} = a_7^{III} \\ a_2^{II} = a_2^I - y_2^0 a_5^I & a_2^{III} = a_2^I - y_3^0 a_5^I & a_5^I = a_5^{II} = a_5^{III} & a_8^I = a_8^{II} = a_8^{III} \\ a_3^{II} = a_3^I - 3y_2^0 a_9^I & a_3^{III} = a_3^I - 3y_3^0 a_9^I & a_6^I = a_6^{II} = a_6^{III} & a_9^I = a_9^{II} = a_9^{III} \end{cases} \tag{4}$$

where the superscripts I, II and III denote the number of the element in the region of the delamination.

The shape functions matrices, the inertia matrices and also the stiffness matrices of elements II and III can be obtained in the same way as it was described for the element I.

DELAMINATED PLATE FINITE ELEMENT: The same concept, as it was for the beam element, is used in case of modeling the delamination region in the composite plate (see Fig.2). The delamination is modeled by three plate finite elements which are connected in the delamination crack tip using the same type boundary conditions. Material

layers in the element are located symmetrically with respect to x–y plane. Each element has eight nodes with five degrees of freedom.

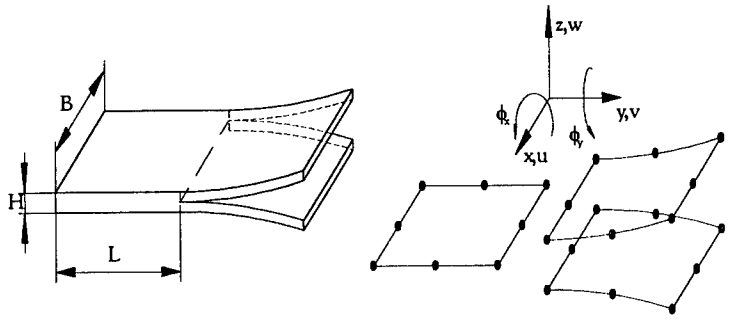


Fig.2. The delamination in a plate modeled by tree finite elements.

Axial displacements u, v and w in a single layer can be expressed as:

$$\begin{cases} u(x,y,z) = u^0(x,y) - z\cdot\phi_x(x,y) \\ v(x,y,z) = v^0(x,y) - z\cdot\phi_y(x,y) \\ w(x,y,z) = w^0(x,y) \end{cases} \tag{5}$$

where $u^0(x,y)$, $v^0(x,y)$, $w^0(x,y)$ denote mid–plane displacements, while $\phi_x(x,y)$ and $\phi_y(x,y)$ denote independent rotations. To approximate the axial mid–plane displacements and rotations biquadratic shape functions for eight–node isoparametric element have been used.

Using standard finite element formulae the inertia matrix of the whole element can be determined as a sum of single layer inertia matrices. Similarly, the stiffness matrix can be evaluated if the strain–displacement matrix for each single layer is previously calculated.

Boundary conditions at the tip of the delamination: To connect elements modeling the delamination area and to satisfy continuity of the displacements the following conditions must be fulfilled (see figure below):

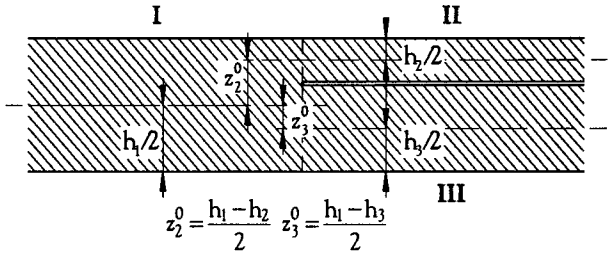


Fig.3. The cross–section of the plate in delamination crack tip.

$$\begin{cases} u_2^0 = u_1^0 - z_2^0 \cdot \phi_x & u_3^0 = u_1^0 - z_3^0 \cdot \phi_x \\ v_2^0 = v_1^0 - z_2^0 \cdot \phi_y & v_3^0 = v_1^0 - z_3^0 \cdot \phi_y \end{cases} \tag{6}$$

where: z_2^0 and z_3^0 are the distances between neutral axes of elements I–II and I–III.

Applying relations (6) to (5), the displacement fields of elements II and III, which model the delamination region, may be evaluated. Analogously, the inertia and stiffness matrices of these elements can be found.

NATURAL FREQUENCIES OF THE COMPOSITE BEAM: Numerical calculations have been carried out for the cantilever beam (see Fig.4.a) of the following dimensions: length 600 mm, height 25 mm and width 50 mm. The beam was made of graphite–epoxy composite (see Appendix). It was assumed that all layers of the beam have the same mechanical properties, i.e. the volume fraction of fibers and the angle of fibers in each layer are identical.

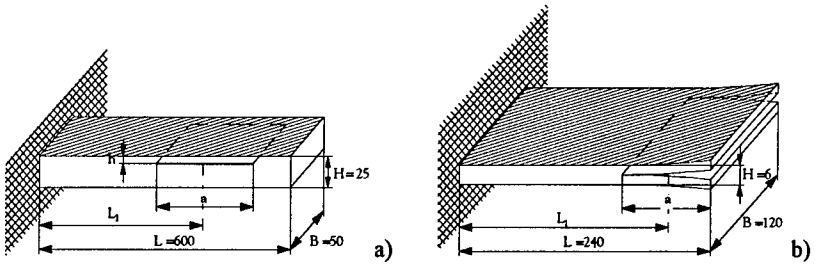


Fig.4. The geometry of the analyzed: a) composite beam, b) composite plate, with the delamination.

The first example illustrates the influence of the delamination position along the beam height on the changes of the first bending natural frequency for four different volumes fraction of fibers. The length of delamination was 37.5 mm (a/L=0.0625) and the center of

the delamination was located 431.25 mm from the free end of the beam (L_1/L=0.71875). The angle of fibers (measured from x–axis of the beam in the x–z plane) was 45°, whereas the relative volume fraction of fibers was: 0.2, 0.4, 0.6 and 0.8. In this case the beam was modeled by 17 finite elements (2 elements in layers modeling delamination and 15 elements outside the delaminated region). The results of numerical calculations are given in Fig.5.a. It is clearly shown that the natural frequency is the most reduced when the delamination is located along the neutral axis of the beam. When the delamination is located near the upper or lower surface of the beam the changes of the natural frequency is negligible.

The next example shows the influence of the location of the delamination along the beam length on the changes of first bending natural frequency - Fig.5.b.. The delamination was located along the neutral axis of the beam. The length of the delamination was 37.5 mm (a/L=0.0625). The same number of elements was used as in the first example. It is easy to find that the changes of natural frequency strongly depend on the location of delamination. For the analyzed beam the largest reduction of natural frequency is observed if the center of the delamination is located at the node of the mode shape associated with this frequency.

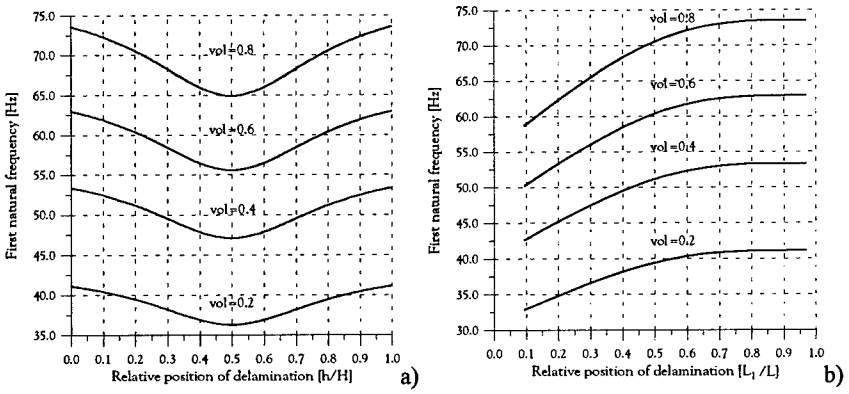


Fig.5. The effect of the delamination location along: a) the beam height, b) the beam length, on the first bending natural frequency.

NATURAL FREQUENCIES OF THE COMPOSITE PLATE: Numerical calculations for the cantilever composite plate have been carried out for the following plate dimensions: length 240 mm, width 120 mm and height 6 mm (see Fig.4.b). The plate was consisted of six layers of materials with changing angle of fibers ±45°. Each layer of the plate was made of graphite–epoxy composite (see Appendix). It was assumed that all mechanical properties except the angle of fibers are the same in each layer. The mechanical properties of the applied material are the same like in case of the beam.

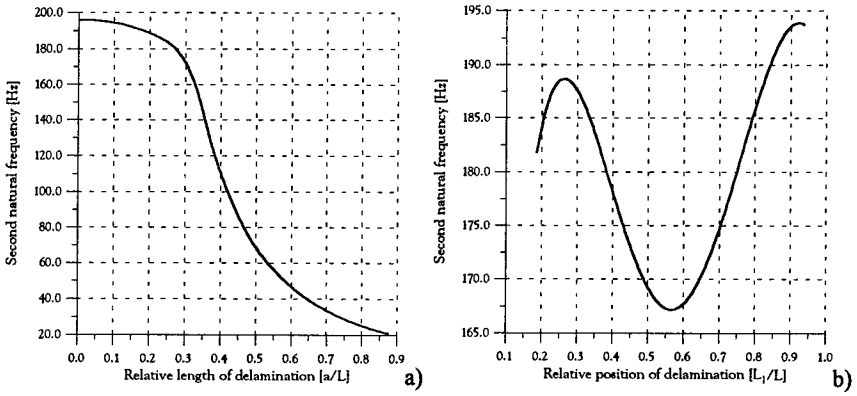


Fig.6. The influence of the delamination: a) length, b) position on the first natural frequency of the cantilever composite plate.

The first example presents the influence of the length of the delamination on the second bending natural frequency (Fig.6.a). It was assumed that the delamination expands from the free end of the plate. The delamination was located along the neutral axis of the plate. It is noted that when the length of the delamination increases the values of natural frequency is greatly reduced.

The last example shows how the location of the delamination along the plate length influences on the drop of the second bending natural frequency (Fig.6.b). The delamination was located along the neutral plane of the plate. The length of delamination was 30 mm (a/L=0.125). It is easy to find that the changes in natural frequency strongly depend on the location of delamination. For the analyzed plate the largest decrease in natural frequency is observed when the center of the delamination is located at the node of mode shape associated with this frequency.

MODE SHAPES OF COMPOSITE PLATE: In this section delamination and orientation of fiber reinforcement influence on mode shapes of a composite plate from the previous section is presented. Calculations were carried out for delamination 60 mm long in the neutral plane of the plate, located at the free end of the plate and extended to its full width. Results of numerical calculations (Figs 7-8) were compared to those obtained for the plate without delamination for two different fiber orientations i.e. 30 and 60 degrees.

From Figs.7-8 it can be seen that changes of analyzed mode shapes depend strongly on associated frequencies and they grow for higher frequencies but some contact effects in the delaminated area for high frequencies may occur (see Fig.7.c,d and Fig.8.c,d). The influence of the orientation of fiber reinforcement is very strong as well.

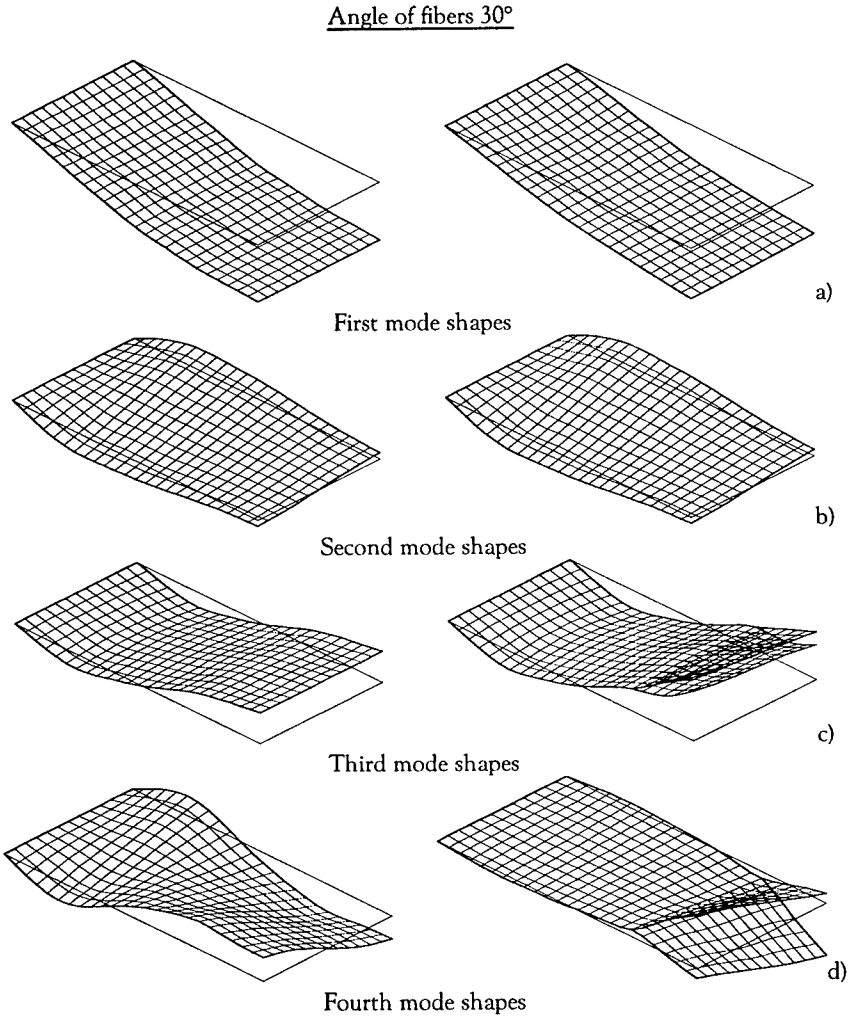


Fig.7. The influence of delamination on the first four mode shapes of the composite plate without delamination (left) and with delamination (right) for the angle of the fiber reinforcement 30 degrees.

It is clearly seen that an appearance of delamination generates new kinds of mode shapes. In those cases mode shapes corresponding to vibrations of the plate without delamination are split into two mode shapes in which one corresponds to vibrations of the delaminated area while the other to the undamaged part of the plate. Also coupling of vibration forms may be observed.

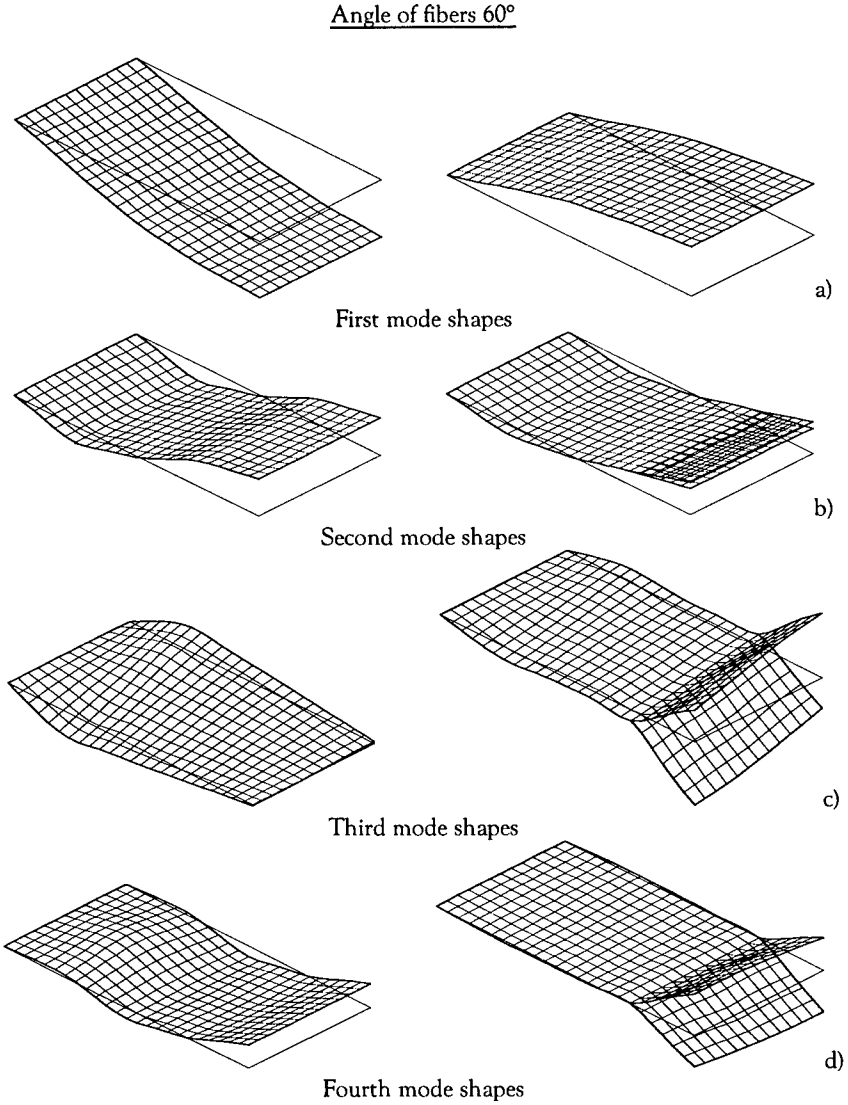


Fig.8. The influence of delamination on the first four mode shapes of the composite plate without delamination (left) and with delamination (right) for the angle of the fiber reinforcement 60 degrees.

Changes of orientation of fibers also influence mode shapes. It can be noticed from Fig.7 and Fig.8 that bending and torsional mode shapes appear in a different order when the angle of fibers changes from 30 to 60 degrees.

CONCLUSIONS: As a result of this work models of the finite beam and plate elements have been elaborated. These elements enable the analysis of the influence of delaminations on the dynamic characteristics of the constructions made of composite materials.

The method of modeling the delamination in composite beams and plates is flexible and allows to analyze the influence of multiple delaminations and also material properties (i.e. position and volume fraction of fibers) on natural frequencies and mode shapes of beams and plates with various boundary conditions.

APPENDIX: The properties of the components of graphite-epoxy reinforced composite [7–8] analyzed in the paper are as follows: • epoxy resin – E_m=3.43 GPa, ν_m=0.35, ρ_m=1250 kg/m^3,• graphite fiber – E_f=275.6 GPa, ν_f=0.2, ρ_f=1900 kg/m^3.

BIBLIOGRAPHY:

[1]D.E.Bently, A.Muszyńska, "Early detection of shaft cracks on fluid–handling machines". Proceedings of the International Symposium on Fluid Machinery Troubleshooting, Anaheim 1986, pp.53–58.

[2]F.D.Ju, M.E.Mimovich, "Experimental diagnosis of fracture damage in structures by the modal frequency method". ASME Journal of Vibration, Acoustics, Stress, and Reliability in Design, 1988, Vol.110, pp.456–463.

[3]W.Ostachowicz, M.Krawczuk, "Dynamic analysis of delaminated composite beam". Machine Vibration, 1994, Vol.3, pp.107–116.

[4]M.Krawczuk, W.Ostachowicz, A.Żak, "Natural frequencies of delaminated composite beam". Proceedings of XII Polish Conference on Computer Methods in Mechanics, Warsaw–Zegrze 1995, pp.171–172.

[5]M.Krawczuk, A.Żak, W.Ostachowicz, "Analysis of natural frequencies of a layered composite plate delamination". Reports of The Fluid Flow Machinery Institute, 1995, No. 307/95, pp.1-78.

[6]S.B.Tessler, S.B.Dong, "On a hierarchy of conforming Timoshenko beam elements". Computer and Structures, 1991, Vol.38, pp.334–344.

[7]A.L.Kalamkarov, "Composite and reinforced elements of construction". 1992, John Willey and Sons, New York.

[8]J.R.Vinson, R.L.Sierakowski, "Conduct of Structures Composed of Composite Materials". 1989, Martinus Nijhoff, Dorchester.

INVESTIGATION OF ELASTIC, PLASTIC, AND CREEP RESPONSE OF A POLYMERIC CLIP USING EXPERIMENTAL AND NUMERICAL TECHNIQUES

Wade D. Bartlett, PE
Mechanical Forensics
PO Box 1958
Dover, NH 03821-1958

James Sherwood, PhD, PE
University of Massachusetts
One University Ave
Lowell, MA

Abstract: The retention force exerted by polymeric surface-mounted cantilever clip retainers was measured during long-term exposure to room temperature, followed by a long-term elevated-temperature cycle. The experimental results were compared to predictions made using finite element analysis (FEA). The FEA was conducted with both a simple Bailey-Norton material model and with a more complicated model incorporating an Arrhenius temperature term. The predicted forces were found to be generally too low at room temperature, and too high at the elevated temperature, as a result of compromises involved in selecting material constants which best fit available tensile and creep material data. The more complicated model offered slightly higher fidelity, but was not deemed significantly better than the simple model, in light of the experimental variability.

Key Words: Creep analysis; Finite element analysis; Polymer;

INTRODUCTION: The electronics industry routinely uses plastic retainers to mount surface components in a variety of assemblies. These retainers must provide sufficient retention force through the lifetime of the unit to hold the retained component. One such polymeric cantilever clip retainer was analyzed using FEA methods, and performance in the elastic, plastic, and creep regimes was measured while exposed to 20°C for 1100 hours, followed by 80°C for an additional 1100 hours. Fifteen exemplar retainers were tested.

The clips were injection molded from the General Electric Plastics (GE) polyetherimide resin ULTEM 1000 [1]. Each clip was approximately 3-cm long and 1-cm wide. A brass insert with a knurled outer diameter and threaded inner diameter was molded into each clip to affix the unit to a substrate. Figure 1 shows a clip installed on a base, prior to insertion of the retained component, which in this case was a transistor. The clips were housed in an oven which was designed to prevent an over-temperature event and for even heating at all clip locations. Custom loadcells allowed periodic measurement of the retention force exerted by each clip.

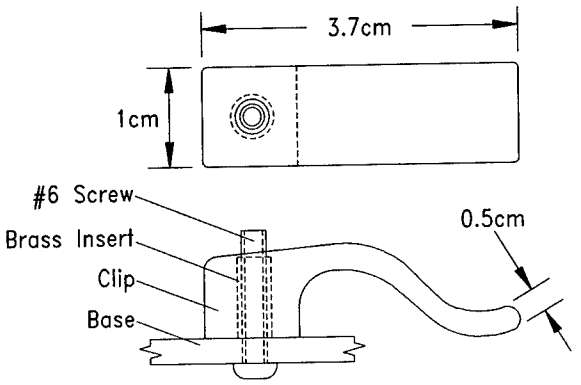


Figure 1: The clip affixed to its base prior to transistor insertion.

A numerical model of the clip was created using the ADINA [2] finite element analysis package. A convergence study was performed to optimize the mesh density. Two material models, one Bailey-Norton type, and one with an additional Arrhenius temperature term were used. Analytical results were compared to experimental results. The relative value of using a more complex material model was evaluated in light of experimental variability.

EXPERIMENTAL PROCEDURE: The test hardware consisted of a calibrated loadcell installed at each clip-location in the oven. Each component was custom designed and fabricated. A standard procedure was used to install each clip so as to minimize variability in load histories.

Ovens: An oven was designed and constructed to accommodate fifteen clips. It consisted of a central aluminum plate measuring 101 x 10 x 1 cm, with 97 x 9 x 8 cm side covers fabricated from 0.076 cm thick steel sheet. An 8-cm diameter hole at each end of the center plate allowed air to circulate through the oven. A motor mounted on one cover drove a fan blade generating a nominal air velocity in the oven of approximately 3 m/s.

The oven was equipped with four 50-Ohm resistors, located in pairs at each end of the oven, to generate heat. Tubular resistors were selected to increase the surface area over which heat transfer could take place. Temperature control was achieved by adjusting the voltage applied to the resistors. The oven was wrapped in approximately 6-cm of fiberglass insulation to reduce energy losses.

Two redundant temperature control systems were installed to reduce the chance of a temperature runaway. Each system utilized an independent power supply and independent small-area fast-response resistive thermal devices (RTDs). The primary controller was set at the target temperature, while the secondary controller, which had override capability,

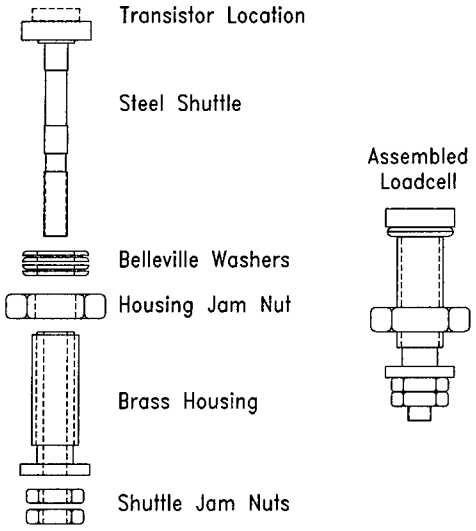


Figure 2: Showing the components of each loadcell and the transistor location.

was set for 0.3°C higher than target. Operation of the secondary system was manually checked periodically.

Loadcells: Each loadcell consisted of three main components: a brass housing, a steel shuttle, and four Belleville washers. Figure 2 shows the components of the loadcell individually and assembled. The housing was threaded into the center plate, and held in place by a single jam nut. The shuttle rode in the housing bore, with the Belleville washers seated between the shuttle and the housing. One washer was nestled into another making a pair. Two pairs were mated with concave sides facing each other, to minimize sliding between the stacks during compression.

After assembly, smooth operation of each loadcell was checked, and each unit was preloaded to between 6 and 12 N by seating the spring-pack and threading two jam-nuts onto the end of the shuttle to compress the springs slightly. Repeatability of the load curves was verified, and the force-displacement curve for each loadcell was documented for calibration.

During testing, the displacement of each shuttle with respect to its housing was monitored, and correlated with these initial calibration curves to determine the retention force being exerted by each clip.

Test Setup Assembly: With the loadcells in place on the center plate, each clip was mounted in a fashion similar to that used on the assembly line. Each mounting screw was tightened to a torque of 110 N-cm, and a transistor was placed under each clip.

Data Collection: The initial retention force exerted by each clip in this study, as well as thirty additional clips, was measured within ten minutes of its transistor's insertion. Once installed, the retention force exerted by each clip was measured at varying intervals, with more frequent measurements taken immediately after the clips experienced a temperature change.

ANALYTICAL PROCEDURE: An exemplar clip was used as the basis for the geometric parameters of the mesh. A mesh convergence study was undertaken to determine an appropriate element density. The model was expanded to include the brass

insert, steel holding screw, and the base to which the clip was affixed. Material constants for the clip were determined using published and unpublished material performance data provided by GE, including stress-strain curves at various temperatures and strain rates, and creep performance data at various temperatures and stress levels.

Clip Geometry: A photograph of an exemplar clip was enlarged for the purpose of determining dimensions of the clip. An outline was defined in terms of a series of straight lines and arcs. Dimensions were verified using a micrometer and an optical comparator.

Mesh Generation: Figure 3 shows the final clip mesh. During the convergence study, the nodes at the base were completely fixed, and a node near the transistor contact area was displaced in thirty steps to a deflection of 5 mm (approximately the same deflection experienced in actual use), over two seconds. The force required to accomplish that displacement was recorded for models with increasingly finer meshes, and was found to decrease monotonically with increasing element density. Increasing the number of elements from 50 to 120 produced 1.4 % change in the predicted force.

By enlarging elements in areas of low stress, and transitioning element sizes smoothly throughout the mesh, a final mesh composed of 105 two-dimensional plane-strain solid elements was created, and was used throughout this study.

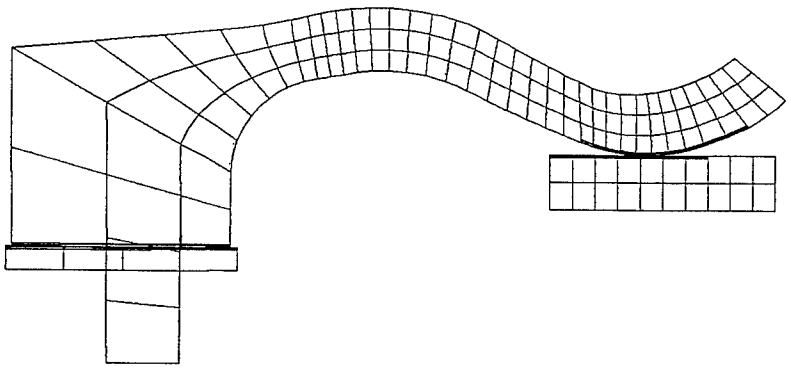


Figure 3: The 105 element mesh used throughout this analysis, showing the clip, base, screw, and transistor.

Elastic / Plastic Model: Preliminary calculations indicated that the maximum strain induced in the clip during normal installation was approximately 8%. This strain was achieved in approximately 2 seconds, during tightening of the base screw. Thus the elastic model was generated based on a strain rate of 0.04 cm/cm/second (1/s). Stress-strain curves for ULTEM 1000 at temperatures of 23, 65, 93, and 105°C were examined. A

bilinear approximation was fitted to the data for each curve. Figure 4 shows one such curve, with the bilinear approximation superimposed on the test data. The slope of the linear segment which intersected the origin represented the Young's modulus. At the yield stress, the slope changed to the secondary modulus, represented by another linear segment. The maximum plastic strain was determined geometrically from the plots.

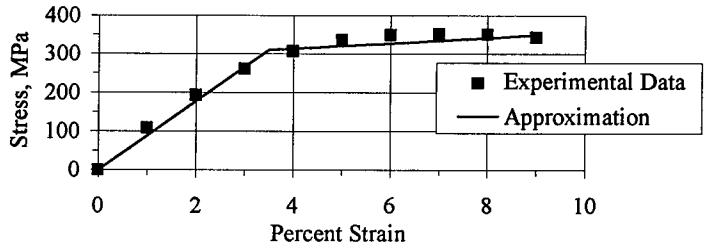


Figure 4: Elastic and plastic performance of ULTEM 1000 at 23°C, showing experimental data and the bilinear approximation used in this analysis.

Creep Model: The GE database included information about the creep performance of ULTEM 1000 measured after 28 hours of exposure to a variety of temperatures and stresses. In additional to that data, GE generated creep data specifically for this project at temperatures and stresses outside the normal design range for this material.

Two material models were implemented: one Bailey-Norton type, and one model which explicitly included temperature in an Arrhenius function, which for the purposes of this paper will be called the Arrhenius model.

The Bailey-Norton model took the form:

$$E_c = A_0 * (s^{A_1}) * (t^{A_2}) \qquad (1)$$

E_c = creep strain (cm/cm)
s = applied stress (N/cm)
t = time (hours)
A_0, A_1, A_2 = material constants

Equation (1) can be rewritten in the y = mx + b form as follows:

$$\ln (E_c / s^{A_1}) = A_2 * \ln (t) + \ln (A_0) \qquad (2)$$

The term on the left hand side can be plotted as a function of ln(t), as shown in Figure 5. The A_1 term was selected to minimize the spread of data sets. A least squares fit line was drawn for each of the data sets, the average of the linear-regression-slopes for all available data sets defined the constant A_2. The average y-intercept value of the regression lines was used as $\ln(A_0)$. Values used in the Bailey-Norton model were $A_0 = 7.38e\text{-}15$, $A_1 = 3.12$, and $A_2 = 0.219$.

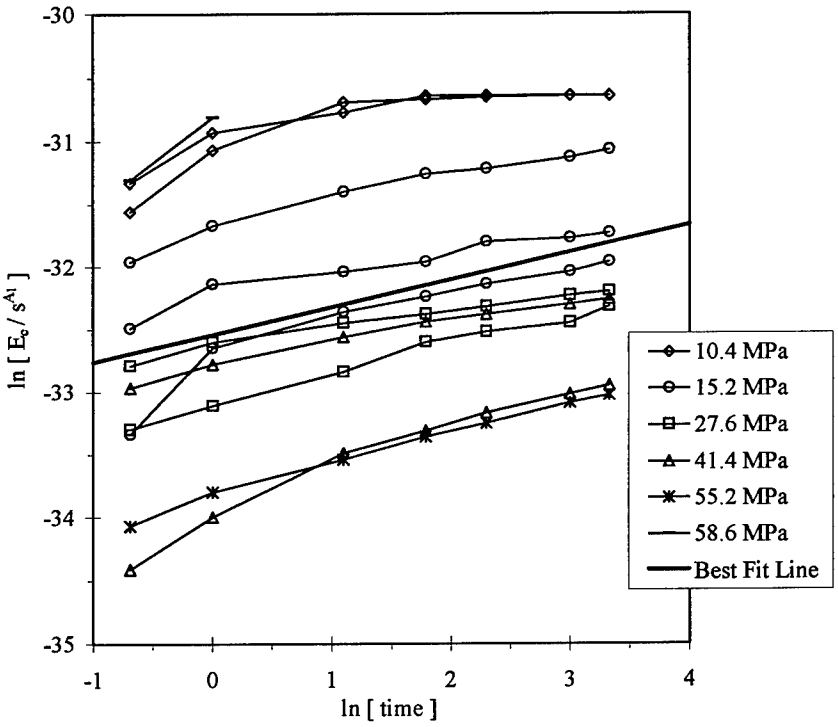


Figure 5: Experimental creep strain data plot used to determine coefficients for the Bailey-Norton material model.

To model the dependence of creep rate on temperature, an additional term was included in the Arrhenius model, which took the form

$$E_c = A_3 * (s^{A_1}) * (t^{A_2}) \, EXP[-A_4 / T] \qquad (3)$$

E_c = creep strain (cm/cm)
s = applied stress (N/cm)
t = time (hours)
T = temperature (K)
A_1, A_2, A_3, A_4 = material constants

The inclusion of the Arrhenius temperature term did not alter the creep rate dependence on stress or time, thus the values of A_1 and A_2 were unchanged from the Bailey-Norton model. To determine A_3 and A_4, the test data were plotted with the $\ln[E_c / (s^{A_1} * t^{A_2})]$ shown as a function of [1/T] in Figure 6. This form of Equation (3) again gives a line of the form y = mx + b, with a slope of A_4, and a y-intercept of $\ln(A_3)$. A least squares-fit line was drawn through all available data to obtain A_1 = 2.1e-11, and A_4 = 2614.

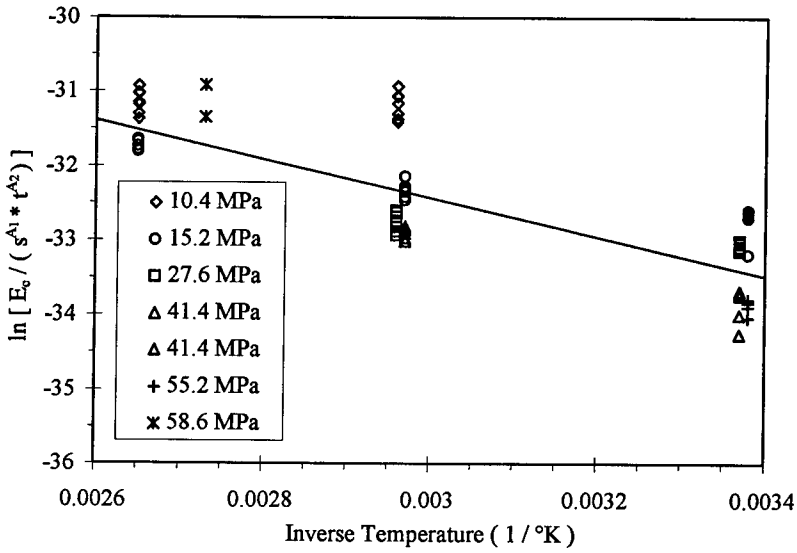


Figure 6: Material data plotted for determination of Arrhenius model coefficients.

ELASTIC / PLASTIC REGIME RESULTS: The elastic-plastic behavior of both material models agreed well with actual test data. Due to the very short time duration of the inital loading (two seconds), the response was primarily controlled by the bilinear stress-strain relationships, which were shared by both models. Thus the variation between the two models was minimal. Figure 7 shows a typical experimental force deflection curve and the predictions made using each material model.

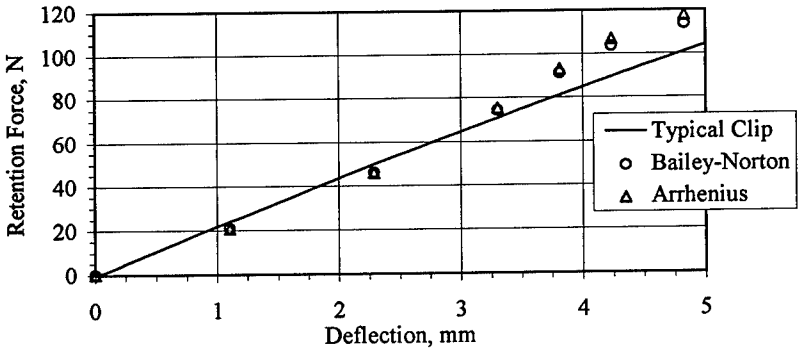


Figure 7: Comparison of elastic-plastic response predicted using two material models with experimental results.

The initial retention force exerted by the 45 clips tested varied from 71 N to 142 N, with an average load of 104.5 N. The bell-curve distribution of the initial retention force is shown in Figure 8.

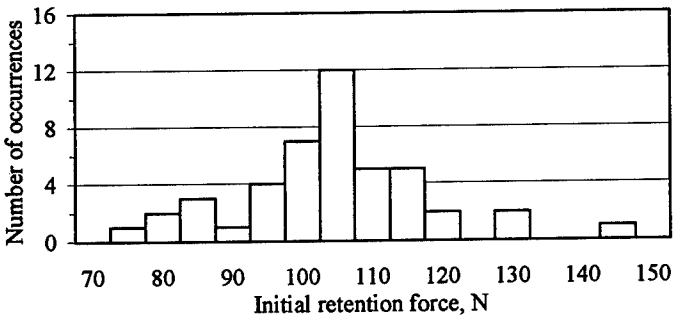


Figure 8: Distribution of initial retention force exerted by 45 clips.

CREEP REGIME RESULTS: Figure 9 shows the experimental minimum and maximum retetion forces as functions of time, along with the analytical predictions. The average retention force exerted by the 15 clips dropped at a decreasing rate, as predicted.

The Bailey-Norton model was found to predict retention forces too low during the room-temperature testing and failed to fully account for the loss in force resulting from the temperature increase. The Arrhenius model predicted retention forces between the experimental minimum and maximum through the room-temperature testing but also failed to capture the result of the temperature increase.

The Bailey-Norton model did not account for the difference in creep response at different temperatures, nor did it provide any means of including the load history. Thus its predictions of temperature effects were based on changes in the elastic/plastic response defined with the temperature-dependent bilinear curves discussed earlier. Furthermore, it predicted forces at each data point as if the clip had been exposed to the current temperature throughout the test. In the present case, this feature manifested itself in the very abrupt change in predicted retention force when the temperature was elevated.

The Arrhenius model, incorporating a temperature-dependent term, predicted a slightly smoother transition, but was still unable to consider load history. It, too, was not able to take full account of the effect of the increased temperature.

After 1500 hours, several of the experimentally investigated clips failed due to stress rupture. This phenomenon was not incorporated into either of the two material models as no information about it was available.

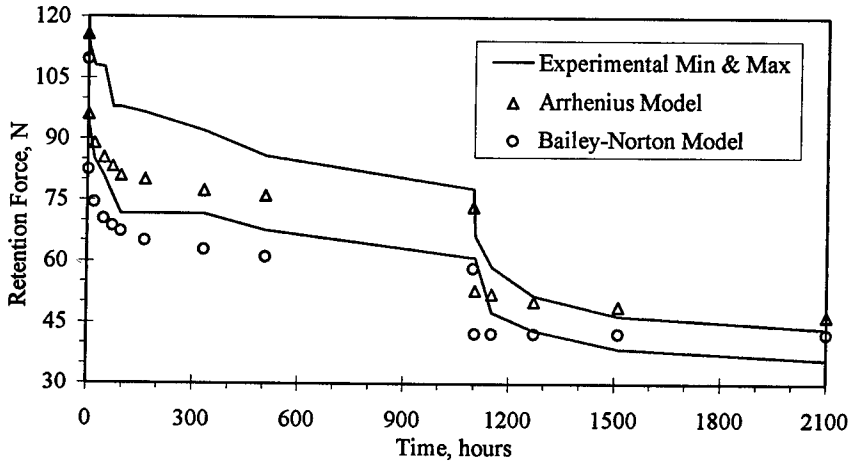


Figure 9: Comparison of experimental high and low data with analytical predictions.

CONCLUSIONS: Both material models predicted the elastic/plastic response within the experimental range. The process of determining proper coefficients for a model incorporating an Arrhenius term was more complicated than that required for the Bailey-Norton model, and failed to provide results which were significantly more accurate in light of the wide experimental data spread.

Most of the available material performance data was generated at temperatures between those of interest in this analysis. This preponderance of intermediate temperature data weighted the material constants in favor of those temperatures. The compromises made in fitting each model to the entire body of test data were evident in both models, which generally predicted holding forces lower than actual at room temperature, and higher than actual at 80°C. It is anticipated that using data generated at the temperatures of interest will produce more robust models.

The inclusion of stress-rupture performance data would be important in any further examination of long-term response, as this was a common failure mode.

ENDNOTES:

[1] ULTEM is a Registered Trademark of General Electric Company
[2] ADINA = Automatic Dynamic Incremental NonLinear Analysis, 71 Elton Avenue, Watertown, MA

EFFECT OF INTERGRANULAR SURFACE ATTACK ON THE FATIGUE PROPERTIES OF AM355 MATERIAL

Victor Champagne
Scott Grendahl
Marc Pepi
US Army Research Laboratory - Materials Directorate
Aberdeen Proving Ground, MD 21005-5059

Dr. Kirit Bhansali
US Army Aviation and Troop Command
4300 Goodfellow Boulevard
St. Louis, Missouri 63210

George Liu
Westar Company
St. Louis, Missouri 63044

Abstract: A critical flight safety component from an Army attack helicopter failed catastrophically in service. The part was fabricated from AM355 material which is a semiaustenitic precipitation hardenable stainless steel. The U.S. Army Research Laboratory and the U.S. Army Aviation and Troop Command performed an in-depth metallurgical analysis of the broken component which revealed that the premature failure was attributable to fatigue. A multitide of fatigue cracks initiated as a result of crevice corrosion which was aggravated by fretting. Further analysis of material taken from various stages of processing revealed an intergranular surface attack which was determined to be caused from acid pickling during primary processing. The intergranular attack was sufficiently deep to entrap hard particles of silicon carbide (SiC). The particles of SiC were remnant from a sanding operation conducted on the surface of the parts. It was hypothesized that the surface intergranular attack may have led to premature crack initiation because of several factors; the breakdown of the protective passive layer, the stress concentration effect of the attack and the abrasive action of the trapped SiC particles within the intergranular surface attack. A study was subsequently conducted to quantify the effects of varying degrees of surface intergranular attack on the fatigue properties of the material. Fatigue specimens were machined from actual components taken from inventroy and from fielded components. The specimens were examined for evidence of intergranular surface attack utilizing light optical microscopy and seperated into four groups which described the degree of severity of attack based on appearance and depth measurement; none, light, moderate, and heavy. Fatigue test data showed a direct relationship between the number of cycles to failure and the severity of the surface intergranular attack. The specimens containing heavy or severe attack had fewer cycles to failure than the remaining specimens. This trend was observed throughout the entire study. Recommendations were made to control the amount of surface intergranular attack or to remove it altogether mechanically or by electrpolishing.

KEYWORDS: Failure Analysis, AM355, Fatigue, Stainless Steel

Background: The U.S. Army Aviation Troop Command (ATCOM) reviewed the manufacturing facilities of an army helicopter subcontractor producing main rotor strap packs, in response to a surface condition observed on the strap pack laminate material. The material, AM355 stainless steel, contained small black spots, which appeared after a pre-stress operation performed on finished strap pack assemblies. ATCOM requested an analysis of these anomalies by the Materials Directorate of the U.S. Army Research Laboratory (ARL). ARL concluded that the black spots were caused by entrapped particles on the surface that fretted during assembly. These particles had been lodged within areas of surface intergranular attack. As a direct result of that analysis,[1] ATCOM requested ARL to investigate the intergranular surface attack that was the precursor and probable root-cause of the black spots. ARL concluded that this condition was inherent to the primary processing of the material by the steel mill and existed on the material before the manufacture of the strap pack laminates. Further investigation revealed that the surface intergranular attack of the AM355 laminate material was most likely the result of prior pickling. To produce sheet material having a thickness of 0.014 inches, a series of itteritive reducing and annealing operations must be performed. Pickling was necessary to remove heat treat scale during this process. Until recently, the intergranular attack was being removed from the parts by a mechanical sanding operation. ARL performed fatigue and tensile testing of the AM355 laminate material to explore the effects of the surface intergranular attack. A detailed test plan was formulated and material received by ARL from previous investigations of the Main Rotor Strap Pack was used. The purpose of this study was to determine if the fatigue properties of AM355 stainless steel are adversly affected by an intergranular surface attack caused by "pickling" during primary material processing.

Figure 1 shows how the strap pack attatches to the main rotor hub assembly. The strap pack assembly consists of a series of 22-24 separate laminates stacked upon each other and bolted together. The main rotor blades are connected to the strap pack assemblies which allow for pitch control and rotation.

Test Plan: Material with varying degress of attack, including heavy, moderate to heavy, moderate, and a final catergory of very light to no attack, was used in the test plan. ARL received three sets of laminates upon which fatigue and tensile tests were to be performed. The material was designated as shown in Table 1. Group A consisted of unprocessed laminates, meaning the material was not reamed, end-milled, or edge broken. In addition, laminates were tested from two strap packs previously received by ARL during another investigation. Strap pack laminates were taken from SP#6800 (Group D), that had 614 prior flight hours, and from SP#7888 (Group E), that had 248 flight hours. Table 1 outlines the test plan designed to assess the effects of the intergranular surface condition on the tensile and fatigue properties of the material. The number of specimens representing each group was limited to the strap pack assemblies available for testing.

[1] Grendahl, S., and V.K. Champagne. "Analysis of AH-64 Apache Strap Pack Laminate Surface Anomalies." U.S. Army Research Laboraty, 26 October 1995.

Table 1. Laminate Material Test Plan

Group	Intergranular Condition	No. of Fatigue Specimens	No. of Tensile Specimens
A	Heavy	2	1
B	Moderate-Heavy	10	3
C	Moderate	16	2
D	Very Light - None	7	1
E	Very Light - None	7	1

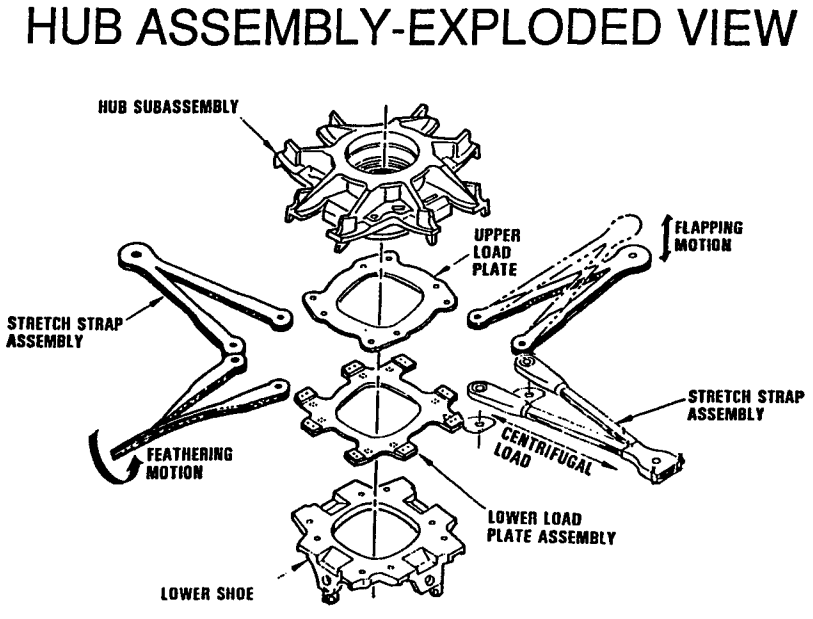


Figure 1. Schematic of main rotor hub assembly showing location of strap pack.

Table 2. Fatigue Test Results

Specimen Designation	Cyclic Stress Amplitude	Cross-Sectional Area	Projected Loads (Pounds)			Cyclic Stress	Cycles To Failure
			Max.	Min.	Mean		
A1	untested	0.0096093					
A2	66.7	0.0097032	1362.0	68.0	715.0	647	53190
A3	57.3	0.0096753	1167.0	58.4	612.8	554.4	3 Million
B1	64.9	0.0092818	1269.0	63.4	665.3	602.8	165352
B2	64.6	0.0093242	1269.0	63.4	665.3	602.8	3 Million
B3	72.2	0.0092995	1413.2	70.7	742.0	671.3	60221
B4	86.9	0.0092592	1649.1	85.9	889.0	804.6	17101
B5	62.7	0.0092860	1225.7	80.9	643.3	582.4	3 Million
B6	77	0.0092812	1505.5	75.1	790.3	715.2	35513
B7	67.5	0.0092459	1314.8	66.0	690.4	624.4	28540
B8	67.5	0.0092555	1314.8	66.0	690.4	624.4	84361
B9	68	0.0092187	1319.2	66.2	692.7	626.5	56138
B10	65.1	0.0092637	1269.0	63.4	665.3	602.8	291466
C1	69.2	0.0093466	1362.0	68.0	715.0	647	44611
C2	57	0.0093014	119.0	56.4	587.2	531	3 Million
C3	64.3	0.0093543	1266.3	63.3	665.0	601.5	70422
C4	64.7	0.0092917	1266.3	63.3	665.0	601.5	3 Million
C5	66.9	0.0093280	1314.3	65.7	690.0	624.6	87941
C6	62	0.0093194	1217.0	60.9	639.0	578	93117
C7	62.1	0.0092993	1217.0	60.9	639.0	578	62477
C8	54.2	0.0092814	1069.9	53.6	561.3	508.2	3 Million
C9	59.7	0.0093014	1168.0	58.3	613.3	555	3 Million
C10	62	0.0093197	1217.0	60.9	639.0	578	137717
C11	79.6	0.0092806	1552.0	77.6	815.0	739.2	31960
C12	59.5	0.0093118	1167.0	58.4	612.8	554.4	155895
C13	82	0.0093489	1595.0	250.0	916.8	???	42218
C14	89.3	0.0093111	1750.7	87.5	919.1	831.6	10986
C15	79.3	0.0093235	1552.0	77.6	815.0	739.2	66804
C16	69.8	0.0092707	1362.0	68.0	715.0	647	77780
D1	67.5	0.0092547	1314.3	65.7	690	624.4	55565
D2	67.4	0.0092598	1314.3	65.7	690	624.4	47986
D3	74.9	0.0092499	1459	73.0	766.0	693	88805
D4	69.8	0.0092726	1362.0	68.0	715.0	647	71278
D5	79.8	0.0092587	1552.0	77.6	814.8	814.8	37649
D6	64.9	0.0092619	1265.4	62.9	664.2	601.3	3 Million
D7	65	0.0092292	1264.0	62.8	663.6	600.6	3 Million
E1	71.9	0.0093102	1408.5	70.4	739.3	669.2	43237
E2	79.5	0.0093004	1552.0	77.6	815.0	739.2	23394
E3	71.5	0.0093704	1409.7	70.5	740.1	669.6	99601
E4	73.8	0.0093831	1459.0	73.0	766.0	693	40925
E5	69.2	0.0093462	1362.0	68.0	715.0	647	3 Million
E6	71.3	0.0093882	1409.7	70.5	740.1	669.6	87210

Table 4. Fatigue Fracture Origin Sites

Specimen	Failure Location
A1	untested
A2	surface
A3	not failed
B1	edge
B2	not failed
B3	edge
B4	edge
B5	not failed
B6	edge
B7	edge
B8	edge
B9	edge
B10	edge
C1	edge
C2	not failed
C3	edge
C4	not failed
C6	edge
C7	edge
C8	not failed
C9	not failed
C10	edge
C11	edge
C12	edge
C13	edge
C14	edge
C15	edge
C16	edge
D3	surface pit
D4,D5	edge
D6,D7	not failed
D1,D2	edge
E1	edge
E2	edge
E3	edge
E4	edge
E5	not failed
E6	edge

Fatigue and Tensile Testing: A 50 kip frame Instron, Model 8502 testing machine with a 5 kip load cell was used for fatigue and tensile testing. The specimens were taken from actual strap pack laminates and were approximately 0.014 inches thick having a rectangular "dog bone" geometry. The length was 8.0 inches and the width was 1.25 inches, with a gage length of 2.50 inches. Each specimen was precisely measured with a Mituyo Mikematis micrometer Model #MK 100E.

The results of the fatigue tests are presented in Table 2 and in graphical format in Figures 2-4. Figure 2 shows the data plotted with "best-fit" approximations according to the equation y=(a+b ln(x) +c/x^2), which was acquired from curve-fitting software. A least-squares regression fit is illustrative in Figure 3 and finally, a power law fit is applied

to the data in Figure 4, for comparative purposes. Data has also been plotted for Kt=1 (smooth surface condition)and Kt>2 (notched specimen).

The results of the tensile testing are presented in Table 3.

Table 3. Tensile Test Results

Specimen Designation	Cross-Sectional Area	UTS (KSI)	0.2% Yield Stress (KSI)	% Elongation
A4	0.0096	240	194	15.6
A5	0.0096	242	227	17.9
B11	0.00927	251.3	219.8	24.1
B12	0.00927	249.7	210.4	24.6
B13	0.0093	254.8	212.1	23.6
C17	0.0093	250	189	19.6
C18	0.0093	246	185	17.4
C19	0.0093	250	179	17.0
C20	0.0093	251	198	19.7
D8	0.0092	253	249	15.4
D9	0.0092	256	250	15.5

Optical and Electron Microscopy: The fracture halves of the fatigue specimens were examined to determine if the point of origin was located at an edge or a surface flaw. The fracture origin sites are listed in Table 4.

Scanning electron microscopy was used to determine the mode of failure and location of the fatigue crack origin(s). The specimens failed predominantly from edge defects caused by sanding which formed scratches perpendicular to the specimen length. When the specimens failed from the edge, these scratches were the origin of fracture, as expected. Figure 5, shows specimen E1 from Strap Pack #7888 with an edge fracture origin (arrow denotes origin) as a result of a scratch from the sanding procedure. Upon closer examination, it was observed that the root of a scratch on the edge of the specimen was the origin of fracture. Specimens from Strap Pack #7888 did not contain any intergranular attack of the surface.

Specimen D3 from Strap Pack #6800 contained a "thumbnail" crack region. The radial lines converge to a point at the surface. Since both Strap Packs #7888 and #6800 had been in service, it was expected that pitting and subsequent surface failures would be observed. Specimen D3 from #6800 failed from a surface pit as depicted in Figure 6. There was no intergranular fracture morphology along the edges of the specimens from Strap Pack #6800. This was expected, since Strap Pack #6800 had minimal to no surface intergranular attack.

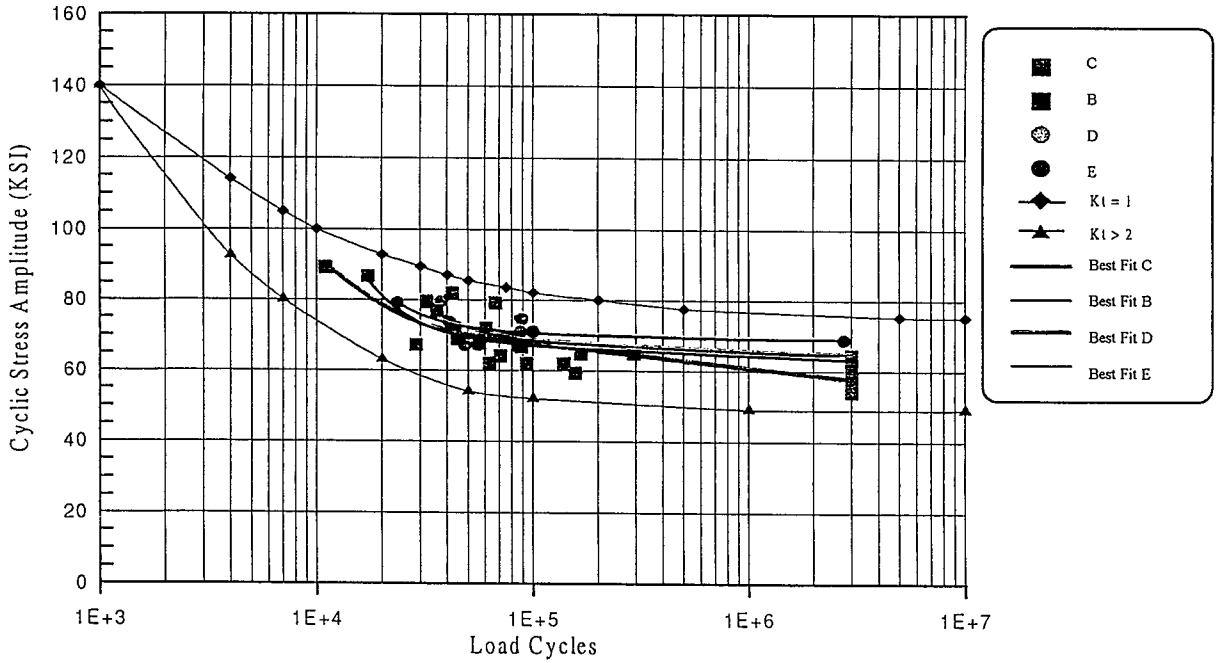


Figure 2 AM355 axial fatigue test - S-N curve; Kt = 1; T = 0.014; G = 0.660; R = 0.05; best-fit curves.

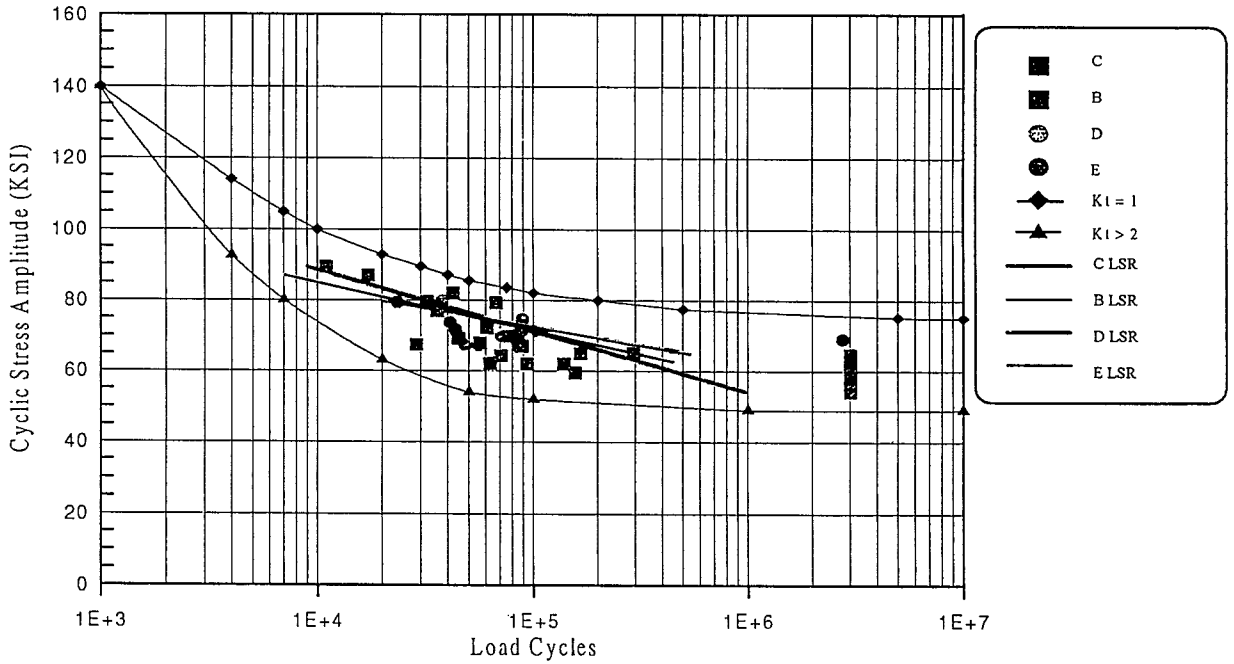


Figure 3 AM355 axial fatigue test - S-N curve; Kt = 1; T = 0.014; G = 0.660; R = 0.05; least-squares regression fit.

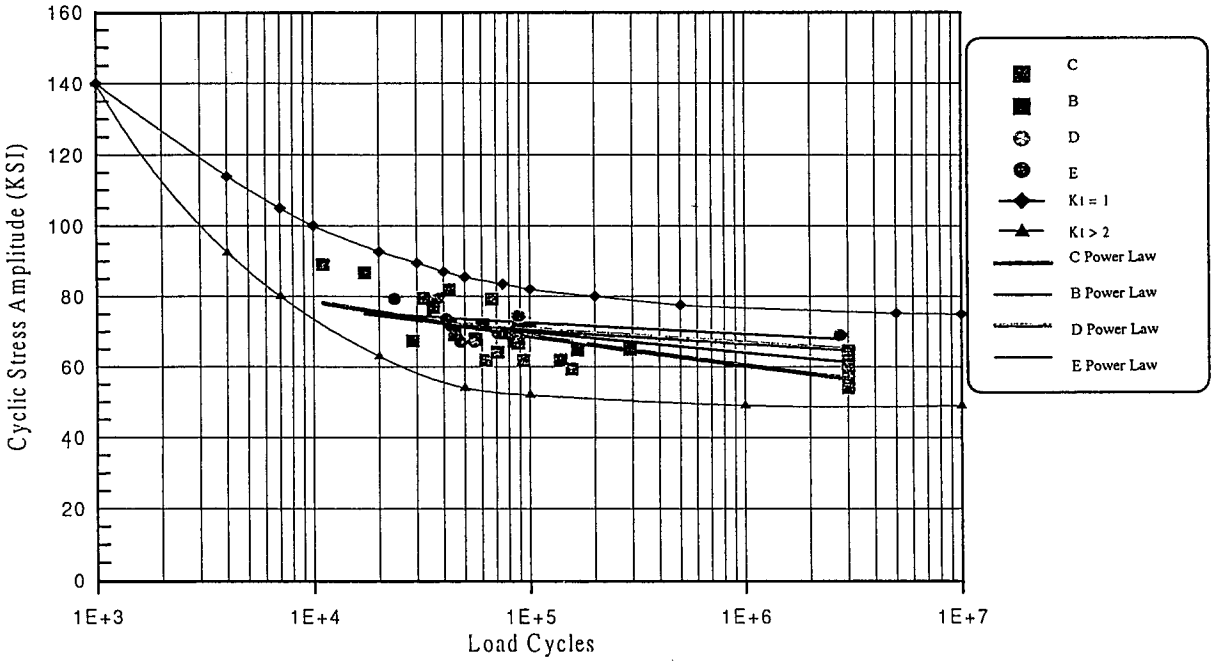


Figure 4 AM355 axial fatigue test - S-N curve; Kt = 1; T = 0.014; G = 0.660; R = 0.05; power law fit.

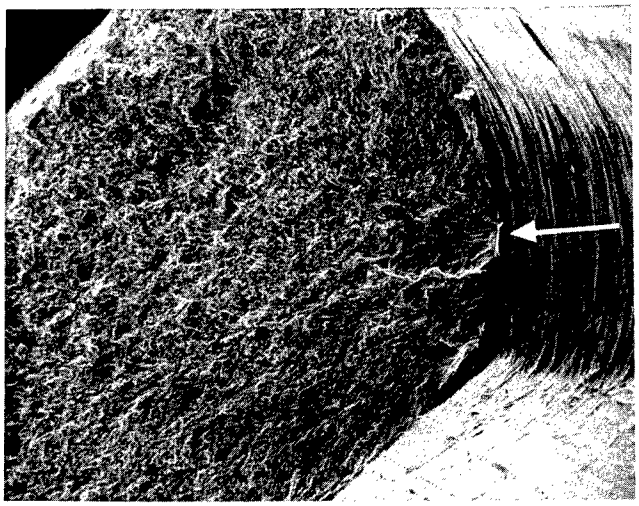


Figure 5 SEM fractograph of E1 showing edge scratches at crack origin. Mag. 250x.

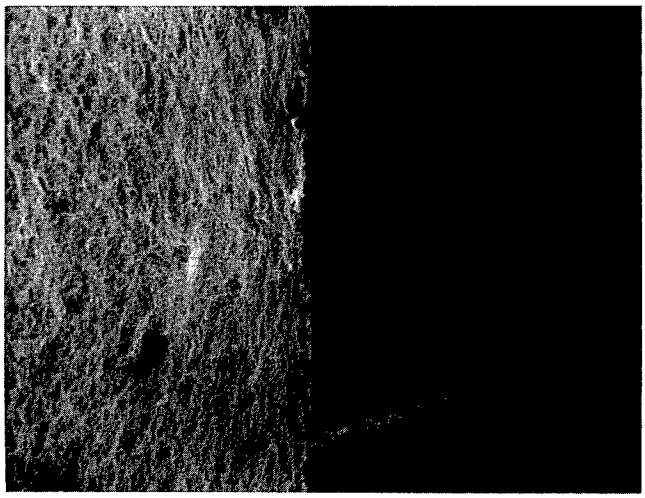


Figure 6 SEM fractograph of D3 containing a corrosion pit at the crack origin. Mag. 250x.

The specimens containing moderate intergranular surface attack (Group C) also failed from surface surface scratches induced during the edge finishing process, shown in Figure 7. However, Group C specimens contained characteristics near the surface not witnessed on the strap packs that had experienced prior service (#7888 and #6800). The intergranular attack of the surface of these specimens allowed the fatigue crack front to progress along an intergranular network near the surface. Figure 7 shows the fatigue crack following the intergranular network near the surface of the specimens(arrows denote intergranular morphology along edges).

Figure 8 shows the edge of specimen C12. The intergranular surface attack allows the crack to progress in an intergranular mode near the surface. It can be observed in Figure 8 that a few grains were removed entirely as the crack progressed around them. There was also significant secondary cracks on the surface of the specimen adjacent to the main crack front.

The group A specimens contained the deepest intergranular surface attack observed. Figure 9 shows specimen A2 which failed from a surface flaw. The proximity of the surface flaw origin to the nearest edge can be seen. Closer examination revealed an intergranular morphology at the origin, shown in Figure 10. Seperated grains and severe secondary cracking are clearly evident as well as intergranular morphology along the edge. The specimens that did not contain this intergranular network did not have the crack front progressing in an intergranular mode near the surface. The crack front progressed by a transgranular mode until transitioning to complete dutility.

Discussion: The optical and electron microscopy results show that the surface intergranular conditon has an effect on the initiation and progression of a fatigue crack front. The intergranular network of attack on the surface of the material has been measured (in previous studies) to be approximately 100-250 μin deep, varying with the location measured and the coil of material from which the specimens were acquired. The intergranular attack is present on both sides of the material and constitutes approximately 2.5% of the total cross-sectional area (using 175 μin as the average depth of intergranular surface attack and the fact that the material is only approximately 0.014 in thick). The intergranular attack lowers the fatigue life of the material simply because a fatigue crack front can easily follow the intergranular network of attack near the surface, and less energy is required to initiate and propogate the crack. The specimens with little to no surface intergranular attack failed from either edge scratches or surface pitsas did those with moderate surface intergranular attack because these defects were deeper than the attack. However, the two specimens of Group A (Heavy Attack) failed from the surface intergranular attack. These specimens had the most severe intergranular attack. Although, only two data points were obtained from Group A, it appears that severe intergranular attack can lower fatigue life when compared to the baseline values (Kt=1). The resultant fatigue crack progressed along the intergranular network near the surface edges of the specimens that exhibited the heavy attack. This mode was not evident on the specimens that did not contain significant surface intergranular attack. However, in the

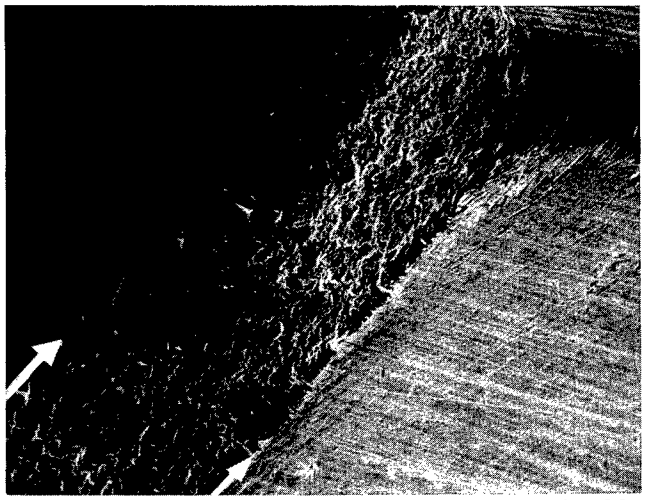


Figure 7 SEM fractograph of C13 showing intergranular fracture at the edge. Mag. 250x.

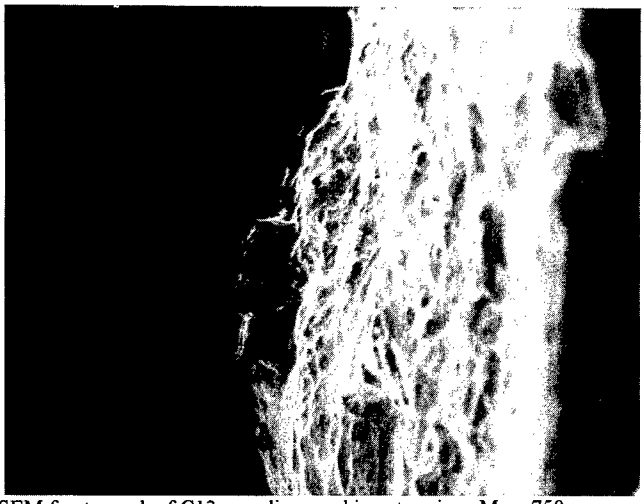


Figure 8 SEM fractograph of C13 revealing cracking at grains. Mag. 750x.

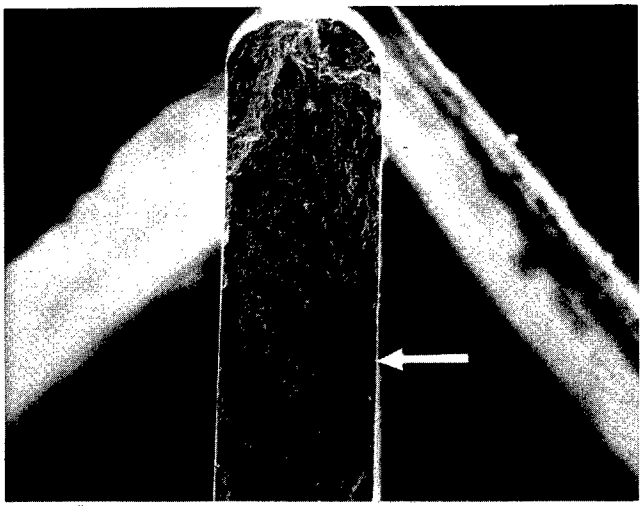


Figure 9 SEM fractograph of A2 showing crack origin at the surface. Mag. 100x.

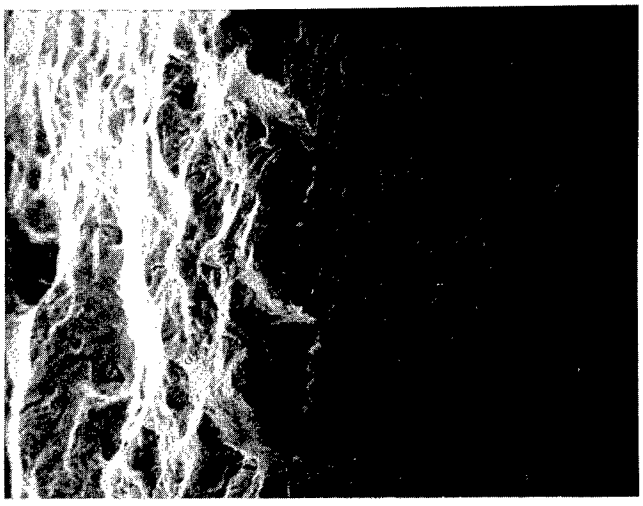


Figure 10 SEM fractograph of A2 showing intergranular surface attack and origin. Mag. 500x.

graphical presentation of the fatigue data, it can clearly be seen in all three fits of the data that the Group B and C specimens also have lower overall and projected fatigue lives. It would be warranted to further study the material with the heaviest intergranular attack (Group A) since it would be expected to have the lowest fatigue life because of the severity of the intergranular attack of the surface. The tensile data all fell within the normal values for the material. It appears that the surface intergranular condition does not have a significant effect on the tensile properties of the material.

Conclusion: The S/N curves generated as a result of fatigue data produced from specimens of AM355 material having intergranular surface attack are lower (2-10 ksi cyclic stress amplitude for similar cycles) than those for similar specimens without this surface attack.

Recommendations: The results of this work demonstrate a reduction in fatigue resistance for specimens of AM355 material having an intergranular network of surface attack. The effect of the intergranular attack on the fatigue properties needs to be quantified by uising material from the same coil (i.e., the same heat treat and surface conditions). Then the material could be tested with the intergranular attack present and with it removed, while all other variables remain constant. Data from a study of this nature could be used to project the fatigue life of material now in service.

MAINTENANCE OF TREADBELT SHOES FOR THE SHUTTLE TRANSPORTER

R. Lee [1], P. Becker [2] and N. Salvail [1]

(1) LO-MSD-2T, Kennedy Space Center, Florida 32899
(2) PZ-D2, Kennedy space Center, Florida 32899

Abstract: The vehicle that transports the space shuttle to the launch pad has 8 tracks, each consisting of 57 shoes. Each shoe is made of AISI 86B30 cast steel and weighs about 0.9 metric ton. So far, twelve shoes from a batch of 228 had a premature failure by cracking at a service mileage of less than 700 miles. In contrast, the shoes of the other batches, with service mileage in excess of 1000 miles, fail primarily by the wear of the pin connection areas. Since these areas can be repaired by weld overlay, cracking is considered especially damaging. A statistical analysis based upon the Weibull distribution indicated an early wear-out of the subject batch. A subsequent investigation revealed that the subject batch had higher hardness and lower fracture toughness than the other batches and that cracks were initiated from the same area. It was also found that the cracking process involved significant plastic deformation and the crack mouth opening was substantial. Since the cracking process was not sudden and brittle, it was concluded that reliability of the vehicle would not be compromised by continued use of the subject batch; cracked shoes could be found and removed before a complete failure occurred. A continuous monitoring routine was established and refurbishment of the worn-out shoes was expedited. Presently, less than 60 shoes from the subject batch remain in service. They are mounted in one track for ease of inspection and maintenance. Failure to provide transportation services for the shuttle launch operation has been successfully prevented by a continuous maintenance plan based upon simple visual and boroscope inspection.

Key Words: Cracking; Fracture toughness; Weibull distribution

Introduction: Once the space shuttle is attached to two booster rockets and the cylindrical oxygen and hydrogen tank, the assembly and the platform are transported to the launch pad on a vehicle known as the Crawler Transporter (CT). The total weight of the space shuttle assembly, the platform, and the vehicle is about 8200 metric tons. The CT travels at the speed of about 1.6 km (1 mile) an hour and it takes about 5 hours to complete the trip between the assembly building and the launch pad. The vehicle consists of 4 double-tracked tractors and is powered by two 2750 hp diesel engines for propelling and two 1065 hp diesel engines for leveling, steering, etc. Each of the 8 tracks has 57 shoes made of boron-containing AISI 86B30 cast steel. Each shoe is about 2.3 m long and 0.6 m wide, and weighs about 0.9 metric ton. Total number of shoes for two CT's and spares are in excess of 1000, purchased from three different sources. They will be called in

this paper G1, G2, and B, denoting "good batch 1," "good batch 2," and "bad batch," respectively. Batch B, 228 in total, is the focus of this paper. So far, twelve shoes of this batch have failed by cracking at a service mileage less than 700 miles. In this paper, a cracking failure is defined as the visual detection of a the crack intersecting the exterior surface. In contrast, the shoes of batches G1 and G2, with service mileage in excess of 1000 miles, were needed to be removed from the vehicles when the pin connection areas were worn beyond tolerance. Since these areas can be repaired by weld overlay for continued use, cracking is considered especially damaging. Several of these cracked shoes are shown in Fig. 1.

Fig. 1 Batch B cracked shoes. The ruler laid on the front shoe is 0.3 m long.

Initially, it was suspected that only those with pre-existing casting defects were subjective to cracking. Fractographic analysis was not conclusive because the fracture surface sustained significant post-fracture mechanical and corrosion-induced damage. This paper describes the investigation to find the cause of the cracking failures and the follow-up maintenance procedure.

Trend analysis: A statistical analysis based upon the Weibull distribution was first performed when the sixth cracking incidence occurred in order to find a trend in the failures of Batch B shoes. The 2-parameter Weibull analysis [1] characterizes the relationship between failure rate and service time by the shape (β) and scaling (α) parameters:

$$f(t) = \alpha\beta t^{\beta-1} \exp[-\alpha t^{\beta}] \quad (1)$$

or

$$F(t) = 1 - \exp[-\alpha t^{\beta}] \quad (2)$$

where f(t) and F(t) are the density and cumulative distribution functions, respectively. As the names imply, the former determines the shape of the curve and the latter describes the spread in the time scale. When the relationship is represented on the Weibull probability paper between cumulative failure and time, the shape parameter conveniently becomes the slope of the line [2]. Figure 2 shows the result with six data points (failures). A slope greater than 5 or 6 would indicate that the devices were well-engineered and are approaching the end of service life; a slope less than 1 would indicate a decreasing failure rate or an actual improvement in reliability as the faulty parts are removed from service by failure. In this analysis, the slope was 2.4. As stated previously, it was initially postulated that only those parts with built-in defects (or casting defects) were failing. However, if this was the case, the slope of the relation should be less than 1. Based upon this analysis, the postulation could be discarded.

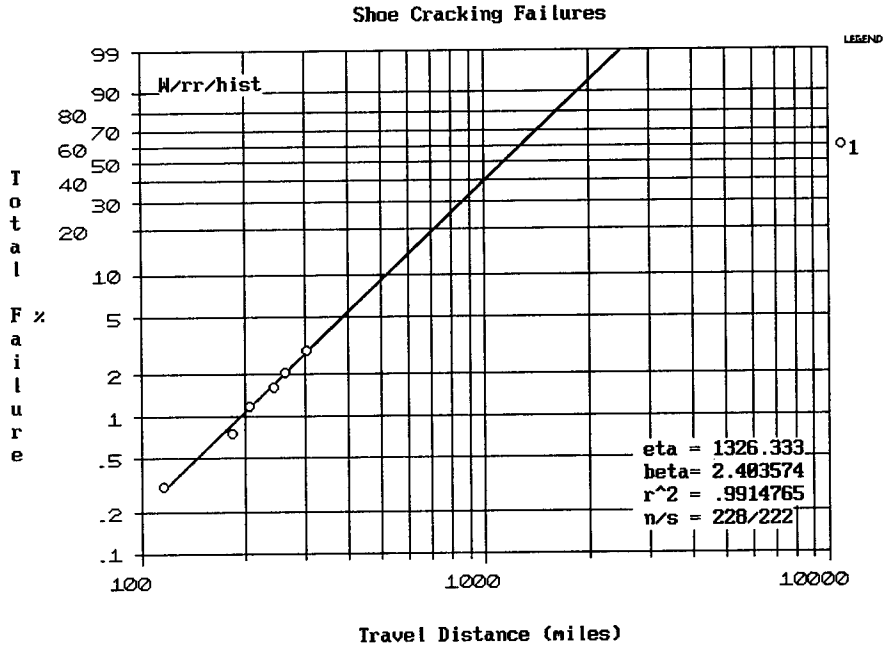


Figure 2. Two-parameter Weibull plot for six failures of Batch B shoes.

Metallurgical analysis: The statistical analysis indicated that the failure mode(s) exhibited by the 6 cracked shoes was most likely generic to Batch B of 228 shoes. An investigation was initiated to find the cause that made the batch more susceptible to cracking. Hardness was measured randomly from all three batches by using a rebounding-ball type tester with

subsequent conversion to the Brinell scale (HB). Batches G1 and G2 showed a range between 240 and 300 HB whereas Batch B had a wide variation between 300 and 430 HB. The required hardness range for these shoes, according to the procurement documents, was between 285 and 331 HB. This was an interesting observation since a large fraction of G1 and G2 shoes were actually lower in hardness than the procurement specification for Batch B. The metallographic microstructure of a cracked Batch B shoe was examined and found to have quenched and tempered characteristics, consistent with the heat treatment requirement. No other metallurgical anomalies were noticed.

In steels, increase in hardness generally increases strength. This, in turn, results in increased resistance to fracture. However, excessive hardness can reverse the trend and reduce the fracture resistance. This property was measured in terms of fracture toughness per ASTM E992, "Determination of Fracture Toughness of Steels using Equivalent Energy Method." . Two shoes from Batch B, previously failed by cracking, and one shoe from G1 were selected. The square compact tension specimens prepared per ASTM E992 had the dimension of 50.8 mm on each side and 12.5 mm thick. Starter cracks were introduced by cyclic loading per ASTM E399. A crack mouth opening of 2.54 mm was found to meet the pre-cracking requirement and used throughout this experiment. Two specimens from Batch B were annealed in a laboratory muffle furnace at 680°C for 2 hours prior to machining in order to reduce hardness comparable to Batches G1 and G2. Table 1 shows the results.

TABLE 1

Comparison of fracture toughness between G1 and B batches.

Specimen No.	1	2	3	4	5
Batch; Shoe No	B; A265	B; A360	G1; 98	B; A265	B; A360
Fracture Toughness ($Mpa \cdot m^{1/2}$)	89	75	167	174	279
Hardness (HB)	342	342	255	285	255
Annealing at 680 °C	N	N	N	Y	Y
Crack Ratio (a/w)	0.448	0.466	0.452	0.454	0.421
Starter Crack (mm)	5.94	6.45	5.99	6.07	4.67

Note: For definition of crack ratio, see ASTM E992.

The results show that the hardness of 342 HB for specimens 1 and 2 was excessive and additional annealing significantly improved the fracture toughness equal to or in excess of specimen 3. The after-anneal hardness numbers for specimens 4 and 5 were different by 30 units although their before-anneal hardness numbers were the same. Since heat treatment effect is greatly enhanced by the minor alloying element of boron in AISI 86B30, it is believed that the concentration of boron was different between specimens 4 and 5.

Two cracked shoes were sectioned to examine the cracks. Cracks started from the mid-length of the hollow channels where tensile stress was the highest and propagated to the

exterior. Extensive plastic deformation was associated with the crack propagation; crack mouth opening was as large as 13 mm. Fracture surfaces were covered with thick corrosion products, indicating that the cracks were developed over a significant length of time. Several Batch B shoes were randomly selected for in-situ boroscope inspection of the hollow channel areas. Several large cracks were noticed but none intersected the exterior, confirming the previous findings that the crack propagation rate was low.

X-ray inspection: Sectioning of two cracked shoes also revealed that the thick sections adjacent to the cracks had a shrinkage cavity. The remaining failed Batch B shoes were examined with X-ray; all were found to contain a large shrinkage cavity at the same location. However, there were no indications that the cavities were contiguous to the cracks. Several Batch G2 shoes (non-cracked) were also X-ray tested. All contained a shrinkage cavity. It was therefore concluded that shrinkage cavities did not necessarily serve as crack initiation sites.

Remedial plan: The results showed that the root cause of the cracking failures was inadequate resistance to fracture associated with excessive hardness and that the resistance could be improved by additional annealing. A decrease in strength was expected by annealing but was not a concern since the other batches (G1 and G2) would have the similar strength. Also, potential crack initiation sites were identified and crack growth rate was found to be low. Based upon these findings, a two-stage remedial plan was drafted. First, inspect the channels with an articulated and illuminated boroscope to identify those shoes that have an active crack. Second, anneal non-cracked shoes to reduce hardness and restore adequate fracture toughness. The vehicle maintenance group, after reviewing this plan to salvage non-cracked shoes, requested the approach to be changed based upon two reasons. First, the as-cast surface in the channels was rough, making the detection of small cracks difficult. In fact, boroscope inspection revealed linear indications in several shoes. It was unclear if the linear indications were shadow lines of surface protrusions and corrosion products, or true cracks. Second, annealing the shoes that contained small but undetected cracks would be a waste of resources. Therefore, the true cost of the proposed plan could not be accurately determined whereas the cost of a refurbished shoe was known. Also, the performance of weld-repaired shoes had been successful whereas the quality of annealed Batch B shoes was unknown. Since the number of worn-out shoes was sufficient to replace all Batch B shoes, it was decided to use the limited resource to repair the worn-out shoes rather than salvaging Batch B shoes with uncertainties.

A continuous monitoring routine was established for Batch B shoes. It involves the simple boroscope inspection of the crack initiation areas and removing cracked ones from service before a complete failure occurs. Most of the Batch B shoes have now been replaced with refurbished shoes. Presently, less than 60 shoes from the subject batch remain in service. They are mounted in one treadbelt for ease of inspection and maintenance. All non-cracked Batch B shoes are kept in a storage area. They may be reutilized in the future when the cost and quality issues associated with annealing can be better quantified. The approach of "replace after cracking but before complete failure" has been able to avoid catastrophic failures and assure availability of space shuttle transport operations.

BIBLIOGRAPHY

1. S. S. Shapiro, "Selection, Fitting, and Testing Statistical Models," in Handbook of Statistical Methods for Engineers and Scientists, H. M. Wadsworth, Ed.,McGraw-Hill, New York, 1990

2 R. B. Abernethy et. al., Weibull Analysis Handbook, AFWAL-TR-83-2079, Wright Patterson AFB, 1983

TAKE THE MYSTERY OUT OF YOUR MOTORS

Adam McMillan
Computational Systems, Inc.
835 Innovation Drive
Knoxville, Tennessee 37932

Abstract: This paper outlines the basic theory of how three-phase AC induction motors are built, work, and fail. With the understanding of these three areas, various methods for monitoring three-phase motors, including flux signature analysis, motor current analysis, and impedance measurements will be explored.

Key Words: AC induction motors; flux signature analysis; impedance measurements; maintenance; motor current analysis

BASIC MOTOR CONSTRUCTION: Three-phase AC induction motors are essentially electromagnets that convert electrical energy into mechanical energy. The most common type of AC motors, polyphase induction motors, are used in a variety of industrial settings, including power generation centers, pulp and paper plants, food and beverage processors, and primary metals production facilities.

The Stator: One of the two critical elements of an AC induction motor, the stator is the stationary component, made up of a core and windings. The core is a slotted steel cylinder made up of thin, laminated sheets. Each sheet is electrically isolated from the others. Laid in the slots of the core are the windings, which consist of coils made from individual strands of copper wire.

In a three-phase motor, the coils are arranged in the stator slots in such a way as to create three phases, or three separate electrical circuits. The circuits are likewise formed into pairs of poles, the number of which will always be a multiple of two. The number of poles determines the synchronous speed of the motor. Synchronous speed is the speed of rotation of the magnetic field.

Stators involve an intricate insulating system, which is vital to motor performance. Each component has its own insulation: the individual laminates of the stator core, the strands of copper wire within the coils, and the coils themselves. An insulating resin is also poured over the entire stator to separate it from ground.

Depending upon their configuration, stators can be categorized into two basic types: random and form wound. Random wound stators contain coils that are made up of numerous turns of wire. Each turn is also made up of individual strands that are formed into loops. The loops are threaded into the slots in the stator, one or two at a time. Because there is no control over how the loops arrange themselves once they are placed in the slots, this type of stator is referred to as random wound. [1]

In a random wound stator, the wire making up the coils has a thin, synthetic film bonded to it, or the wires may be coated with a thicker glass fiber. Both the film and the fiber serve as an absorbent wicking and to bond the coils into a tighter mass. Such coatings also work to enhance the physical distance between conductors in adjacent positions.

Additional insulating systems in the random-wound configuration include the slot cell or cell liner, the center spacer, and the top wedge or "stick."[2] The slot cell, also known as the cell liner, serves as a barrier between the coils and the sharp surfaces of the stator laminates, which could penetrate the coverings on the wires. It also serves as insulation between the core and the coil conductors. The center spacer works much as its name implies--it separates the top and bottom coil sides in each slot. Finally, the top wedge acts primarily as a support mechanism. Acting as a wedge, it keeps the coil securely confined in the slot.[3]

A form-wound motor uses the same basic insulating system, but in a different way. Unlike the random-wound motor, the shape of the coil is predetermined. The coil, which is fully insulated, is then placed into the slots as a whole, rather than threaded in wire by wire as in the random-wound configuration. Another difference between random- and form- wound motors is in the make-up of the coil conductors themselves. Instead of the round, film-covered magnetic wire that is utilized in the random-wound motor, the form-wound machine employs rectangular copper ribbon with a variety of insulating possibilities.[4]

The copper ribbon may be coated with film, but this type of insulation is prone to defects. More commonly, the ribbon is covered with two layers of a polyester/glass fiber. Once the coil is wound and shaped, an insulating material is wrapped around the entire configuration. Special care must be taken when applying the wrapping to ensure thorough coverage and to avoid irregularities.[5]

Both random- and form-wound motors have a ground wall insulation which is the final insulation that is applied. The ground wall is typically a resin that is coated on the windings, then vacuum pressure impregnated and finally baked or cured in an oven.

The Rotor: The second critical component of an AC induction motor is the rotor. The rotor is the rotating component of the motor, which is suspended within the stator core in such a way that it may rotate freely between the poles. There are two general types of rotors, the squirrel cage type and the wound rotor. Unless specified otherwise, the discussions in this paper relate to motors with squirrel-cage type of rotor.

Consisting of heavy copper or aluminum bars, squirrel-cage rotors are used in situations where constant speed is required. There are two basic methods for constructing this type of rotor. The bars may either be short-circuited by end rings (using copper bars), or the bars and the end rings may be part of a one-piece construction, or casting (poured aluminum). Because of the rugged construction the squirrel-cage rotor is widely used throughout industry. The lack of moving electrical contacts make them especially

desirable in environments that contain flammable materials. 6

The bars within the squirrel-cage type of rotor may have different shapes which produce a variety of torque characteristics. Squirrel-cage induction motors have a nameplate which bears a design code letter A, B, C, or D. Established by the National Electrical Manufacturers Association, the design codes represent differences in torque characteristics. Torque is the resistance to rotation due to a force. Within each letter designation, values are tabulated for locked rotor or starting torque, breakdown torque, pull up torque, and slip.[7]

The second basic type of rotor is the wound rotor. Wound rotors are used when it is necessary to control motor speed, or when a higher starting torque is needed. This type of rotor employs a winding similar to that of the stator, with the same number of poles. To allow external resistance to be introduced, the winding ends are connected to slip rings. [8]

BASIC MOTOR OPERATION: When a three-phase AC induction motor is started, the rush of current can be as much as ten times its normal operating current. This current pulses along a path between the three phases in the stator. When that winding is connected to a three-phase electrical source, a separate current flows in each of the three circuits or phases. Between each of these phase, there will be displacement of exactly 120 degrees. This displacement means that phase A will peak 120 degrees before phase B, and that phase B will reach its maximum 120 degrees before phase C. [9]

As the current pulses between the three phases, it produces magnetic flux, which cuts the air gap between the rotor and the stator. As a result of the magnetic flux, current is induced in the rotor. A magnetic field is consequently created around the rotor. This field then works to achieve magnetic alignment with the rotating field in the stator. As the north pole of the stator passes a rotor bar, it induces current along that bar. The south pole of the stator is simultaneously passing a rotor bar on the opposite side of the rotor, inducing current in the opposite direction, resulting in a circular flow of current, turning the rotor into an electromagnet.[10]

As a result of the induced current, the rotor begins to rotate. It is important to note that the revolution of the rotor never catches up to the rotating magnetic field in the stator. Because power can only be developed when the rotor conductors are cut by the rotating field, the rotor and the magnetic field can never rotate at the same speed. [11] The resulting difference between the rotating speed of the magnetic field and that of the rotor is called slip.

The current in the rotor is induced electromagnetically, thus requiring relative motion between the flux and the conductors. If the mechanical load increases, therefore, the rotor slows down. This reduced speed results in a greater relative motion between the flux and the rotor, along with greater voltage and current, as well as more power to handle the increased load. Whenever there is an increase or decrease in the load, the power in the rotor will automatically adjust to compensate for the change. [12]

In contrast to the AC induction motor is the synchronous motor. In this type of motor, a fixed pole magnet is placed inside the stator's magnetic field, and will rotate at exactly the same speed as that field. As its name implies, the synchronous motor operates at a continuous average constant speed, no matter what the load. There are two factors that determine the speed: the frequency of the supply source and the number of poles.[13] The speed is in direct proportion to the frequency and is inversely proportional to the number of poles. For example, a 2-pole motor will run at 3600 rpm, a 4-pole motor will run at 1800 rpm, and a 6-pole motor will run at 1200 rpm.[14]
Synchronous motors are used in high horsepower applications where constant speed, efficiency and power factor correction are critical, such as centrifugal and axial compressors, pumps and fans.[15]

FAILURE MECHANISMS WITHIN MOTORS: The majority of motor failures are bearing-related. There are a number of factors that affect bearing life and performance, including operating environment, operating speed, load, lubrication, or improper or inadequate lubrication. In addition to bearing fatigue, other mechanical faults within motors include imbalance, misalignment, and looseness.

Following close behind bearing faults as the leading cause of motor failures are problems that originate in the stator. When discussing stator failure, it is important to consider the insulating systems described above, and the effect on those systems of the motor's number one enemy--heat. Excessive heat causes those insulating systems to break down, which results in shorting within the windings of the stator. The most common type of shorting, turn-to-turn shorting, occurs when two or more turns of wire in a coil short together. When this short occurs, all of the turns between the two that are shorted are lost. Therefore, if turn one is shorted to turn five, turns two through four are lost, meaning a much stronger current flows between one and five. The heat is also greater than normal, leading to degradation in the winding. This deterioration progresses until ultimately a fault occurs in the ground wall insulation.

Although basic motor operation is universal, each motor is unique. Performance characteristics and failure mechanisms may be vastly different, depending on the application and the operating environment. A wide array of technologies is available for monitoring both stator and rotors, accounting for the divergence in failure mechanisms.

ADVANCED TECHNOLOGIES FOR ELECTRIC MOTOR MONITORING: A typical industrial plant may have hundreds of motors, and determining which of these motors to monitor on a regular basis involves many factors and requires much consideration. To determine which motors should be monitored, one should first consider safety criticality, followed by process criticality, and finally cost criticality. Safety-critical motors are those that would pose an extreme environmental risk and a safety hazard to employees and the general public if they failed. In turn, the failure of process-critical machines would cause the entire process to be shut down. Cost-critical machines are those that would cost a great deal to repair or replace. Motors should be rated for levels of criticality in each category. Using a matrix, machines that are deemed the most critical

receive the most attention, and those deemed least critical may ultimately not even be monitored.

After it has been determined which motors should be monitored, the decision must be made as to how to monitor them. As noted above, most motor failures are bearing-related. These faults can be very adequately addressed with a comprehensive vibration analysis program, which can also detect other mechanical faults, such as imbalance, misalignment, and looseness.

Vibration analysis is conducted using a spectrum analyzer and a software program. Alarm levels for each machine are set, and these levels are used when data is collected. If a machine is in alarm, the maintenance technician knows that an in-depth analysis of the data is needed to determine the cause for the alarm. Vibration analysis facilitates predictive maintenance and allows advanced planning for repairs. It can also indicate the severity of the problem, and whether it is mechanical or electrical in nature.

There are several advanced technologies available today for analyzing problems that are electrical. These technologies consist of both off-line and on-line testing methods. Off-line methods include megger, surge, hi pot, and impedance testing. On-line methods include flux analysis and motor current analysis.

A megger test is used to measure the insulation resistance to ground of the motor, which is the ratio of the DC voltage applied between the winding copper and ground, to the resultant current. Total leakage current to ground, as this resultant current is called, indicates the amount of contamination on the exposed insulation service of the winding, most significantly where there are cracks or fissures.

Weak points in the turn insulation can be discovered through surge testing. The surge test applies a high voltage AC test signal (up to 2 times line voltage plus 1000 volts) for a very short time. A sign wave signature is recorded for each phase and then the three phases are compared to determine the condition of the winding insulation. While surge testing is an effective mechanism for determining the condition of turn insulation, and can be used periodically for checking winding insulation, there are limitations to this method of testing. The surge test, for example, can be destructive, causing failure in weak insulation. The test also requires a fair amount of expertise to conduct, and exposes personnel to high voltage levels. As well, test signals are subject to line interference (induced voltage interference from neighboring circuits that are energized) if conducted from a motor control center, or interference from the rotor.

The hi pot (or high potential) test works on the theory that weakened insulation will fail if it is subjected to a high enough voltage, and that good insulation will remain intact. The purpose of the test is to determine which insulating systems can be expected to fail within a short period of time. Because the hi pot test is potentially destructive, it is rarely used on large generator or motor stators, except when they are being commissioned. Due to this potential for destruction, spare parts should be on-hand or readily available before this

test is conducted.[17]

Off-line tests called impedance measurements can also be used to detect stator defects. Using a Kelvin type measurement, complex impedence measurements are taken on each phase of a three phase motor, testing from the motor control center. The measurements are compared phase-to-phase for balance and can indicate power circuit problems such as loose or dirty contacts as well as motor stator defects. This type of measurement requires that the motor be shut down, and therefore is not as efficient for routine monitoring. It does, however, offer the benefit of providing complete stator diagnostic capability without posing a destructive threat to the motor. Additionally, test signals can be subject to interference from live power circuits laid in parallel to the circuits being tested, and are also subject to rotor interference.

When choosing a technology to monitor AC induction motors, one must keep in mind the overall impact on the operating unit as well as the entire plant. For the routine monitoring of most motors, the logical choice is an on-line test called flux analysis. When an electrical problem is developing, the magnetic field changes, and it is this change that flux analysis detects. Data can be collected in a safe manner, while the motor remains fully operational. A rigid coil is placed on the fan shroud of the motor, so no access to live power leads is required for routine monitoring. Such access to live leads is necessary only in establishing the initial base-line current, making flux analysis the front-runner in terms of safety and convenience. It is important to note that flux can not only detect problems in the stator, but those in the rotor as well.

Before flux analysis, the only way to measure shorts in the stator was to either perform a surge test or measure the low resistance between the phases. Both of these tests, however, require that the motor be taken out of operation for analysis. A megger test can detect short to ground, but only after turn-to-turn or phase-to-phase shorting has already occurred, meaning the motor is already virtually inoperable.

Although flux analysis is the most appropriate technology for routine monitoring of AC induction motors, there are caveats. Flux, for example, can not pinpoint the exact nature or the severity of the problems it detects. When a problem is detected with flux, another technology can be used to diagnose the exact cause. One such technology for detecting rotor deterioration is motor current analysis.

If an electric motor is running properly, it will generate a well-patterned current spectrum. Motor current analysis is a diagnostic tool that uses the electrical characteristics of such spectra to detect problems within the rotor. Motor current analysis is a valuable tool when used in conjunction with flux analysis as a corroborative technology. Its usefulness, however, lies in analyzing problems that have already been detected. Because it requires access to live power leads and requires the assistance of an electrician in most plants, motor current analysis is not recommended for the routine data collection that is necessary to catch emerging problems early, before catastrophic damage occurs.

Temperature measurements are yet another way to monitor motors for electrical problems. Excess heat can be caused by a number of factors, including overloading, clogged cooling fins, high ambient temperature, and problems in the electrical system supplying power to the motor. A common problem associated with monitoring motor temperature has been environmental factors which were not accounted for. Temperature measurements, therefore, could not be made in a reliably trendable fashion. New methods have been recently developed that allow repeatable motor measurements. These measurements can be trended in order to detect a rise in motor temperature, regardless of loading or ambient temperature. When temperature rises, steps can be taken to eliminate the cause.

A PROGRAMMATIC APPROACH: Each of the advanced motor-monitoring technologies discussed have limitations on their applications or barriers to their implementation, which can be avoided or diminished by taking a multitechnology approach. There is no single approach that can account for all of the failure mechanisms that can develop within a motor. Plants wishing to develop a comprehensive program for monitoring motors should therefore consider implementing a system in which several technologies work together to provide a complete picture of motor condition. Although not available currently from a single vendor, a system that utilizes flux analysis, motor current analysis, and some form of off-line measurements can be attained at a moderate cost. The initial investment will be returned many times over through cost-savings obtained from preventing motor failure.

MFPT

DIAGNOSTICS AND PROGNOSTICS I

CoChairs: Howard A. Gaberson
NAVFAC Engineering Service Center

Reginald G. Mitchiner
Virginia Polytechnic Institute & State University

TURBOMACHINERY DIAGNOSIS AND PROGNOSIS BASED ON ACTIVELY SOLICITED IMPULSE RESPONSE

by William D. Marscher and Chang-Wei Jen
Mechanical Solutions, Incorporated
9 Sylvan Way, Suite 360
Parsippany NJ 07054

Abstract: This paper proposes a novel CBM approach capable of reliable fault identification, diagnosis, and prognosis in rotating machinery. The approach supplements the original vibration, which is caused by the machine's operation, with a low level impulse induced by a small exciter built into or attached to the machine. The principle by which the impulse is able to obtain information concerning the machine internal components is similar to the action of sonar, because it relies on the feedback from the exciter signal after it travels through the machine and returns. Unfortunately, this process takes place in a mechanically "noisy" environment that nearly overwhelms the impulse return signal. Therefore, the returning structure-borne sound signal is time-averaged to filter out nearly everything except items that correlate with the randomly applied impulses. A Fourier transform of the time-averaged result produces clean frequency spectra containing natural frequency peaks. From the shape of these peaks, individual component dynamic characteristics can be determined, operational forces can be inferred, and a great many issues can be evaluated which are key to machinery health assessment and prognosis. Such issues include bearing and seal stiffness and implied wear, and existence of cracks in shafts or their supporting structures.

Key Words: Condition monitoring; condition-based maintenance; experimental modal analysis; predictive maintenance; turbomachinery; vibration analysis; vibration sensors; vibration testing

INTRODUCTION: Existing preventive maintenance (P/M) and condition-based maintenance (CBM) systems for rotating machinery are limited in their capability for accurate diagnosis and prognosis. Although most engineers agree that such systems have reduced warranty and maintenance costs in both commercial and military applications, existing methods have not been able to prevent the loss of many billions of dollars of production and unplanned maintenance activity each year. In commercial applications, the flip-side of the burden of these costs is the significant opportunity for increased competitiveness if companies are able to reduce these costs relative to those borne by their competition. In military applications, lack of reliable CBM systems places the effectiveness and safety of missions at continuing risk of being compromised during critical moments, and reduction of this risk increases the effectiveness of our armed forces. Obviously, there is a growing demand among both commercial

and military users of rotating machinery for greatly improved CBM systems that provide superior reliability risk management and control.

SPECIFIC TECHNICAL NEEDS: Most commercial production plants and military services currently use preventive maintenance (replace/repair per pre-determined schedule), and would like to move toward Condition-Based Maintenance (replace/repair only for cause, prior to costly failures). The rate at which this transition can take place is slowed by the lack of reliability of existing CBM systems. Actually, the hardware in modern CBM systems has the potential for sound diagnostics and prognostics. Development of improved methodology more than improved hardware is required to provide the step-change in reliability which is required for this transition to become a sensible one in most applications.

Although different sensors and approaches are needed for different rotating machinery applications, what is common to most applications is that the CBM equipment must be inexpensive, easily installed, user friendly, and require a minimum commitment for installation and upkeep by overburdened maintenance groups. Artificial intelligence including first-line decision making would be strong benefits, if the information provided can be made reliable enough to avoid the need for human "common sense" and judgment in the great majority of cases. Unfortunately, the reliability of existing vibration-based CBM systems is limited because they passively depend only upon dynamic response of a subject machine to operational forces. Using this information alone, CBM cannot reliably determine causes of anomalies. A system is needed that produces reliable CBM input information, providing an alert for a broad spectrum of problems with a minimum number of false alarms.

DESCRIPTION OF THE NEW APPROACH:

Assessment of the Currently Available Systems: Although a number of health management systems exist which can do a fair job of identifying whether a turbomachine is experiencing problems and needs repair, in the opinion of the authors none yet performs an acceptable job of identifying specific components involved in the problem, or gives reliable estimates of remaining time-to-failure. This is somewhat surprising, because vibration-based predictive maintenance (PdM) and CBM systems are not new concepts, and today are available in many varieties from vibration instrumentation companies. However, there is a common flaw in all existing vibration-based machinery CBM systems. In nearly all cases, they rely exclusively on observing the dynamic response of the system to operational forces produced naturally from some component within the system.

The issue is not whether observation of vibration response to natural operational forces is useful, but whether this is a sufficient indicator of potential machinery system problems. When vibration is significantly higher than normal, particularly in certain frequency ranges, it can be an indicator of mechanical or fluid flow problems within the system, particularly if the vibration increase was a continuing trend over a reasonable period of time. Existing vibration condition monitoring

and CBM systems therefore check vibration levels in discrete frequency ranges by transforming the raw vibration sensor signal in terms of vibration level versus time into a new plot, or "domain", of vibration versus frequency. This is done using a mathematical technique known as the Fourier Transform, usually performed very quickly and automatically by computer through exercising its revised form, the "Fast Fourier Transform", or "FFT", as in the example provided in Figure 1. The resulting vibration versus frequency plot is known as a frequency "spectrum", where the heights of the peaks located at specific frequencies indicate the level of vibration associated with those frequencies.

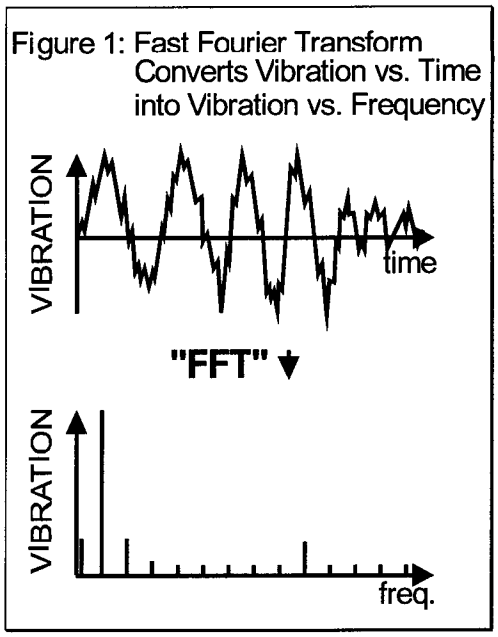


Figure 1: Fast Fourier Transform Converts Vibration vs. Time into Vibration vs. Frequency

Unfortunately, spectral information provided by an FFT analyzer concerning vibration versus frequency due only to natural excitation is necessary but generally not sufficient for determining the specific reasons why a system is not functioning as expected, or how long a system may continue to be operated before catastrophic failure. For example, it could indicate frequencies at which unexpectedly high vibration was occurring, but even so would leave the operator to guess whether the high vibration was caused by excessive fluid forces, a weakened bearing support, resonance of a rotor natural frequency, or other causes. Without resolving the specific reasons for anomalous behavior, any prediction of time remaining until failure is insufficiently informed and unreliable, and it is unclear which components will require repair during future maintenance, thereby requiring that the machine be out of service until the faulty component is located by disassembly and inspection.

The authors propose a fundamentally different approach to vibration-based CBM of military and commercial machinery. The goal of the new approach is to make machinery diagnostics and prognostics more accurate, precise, and reliable. The new approach accomplishes this in part by complementing the amount of experimental data available concerning the system, such that enough data exists to determine the dynamic characteristics of all components critical to life evaluation. Life degradation models of specific components and of the overall system can then be exercised to predict the time until maintenance will be required.

Overview of the Proposed New Method: The authors' proposed CBM approach is capable of reliable fault identification, diagnosis, and prognosis for all types of rotating machinery. The crux of the proposed approach is that it supplements operationally-excited vibration with a very low level impulse response induced by a small on-board exciter. The impulse travels throughout the machinery internal and support structure, and "echoes" back, like a radar signal. The resulting time-series data of the incoming structure-borne sound (unfortunately always far less than the machinery operational noise, even in "quiet" naval applications) is time-averaged in order to enhance and emphasize it, so that it can be reliably determined and evaluated. The time averaging filters the initially 'dirty" signal by checking to see which part of the total received signal correlates with the occurrence of the randomly applied impulses, and statistically de-emphasizing any part of the signal which does not correlate with the "pings" or impulses. This correlated response emphasizes the impulse response, which produces frequency spectra containing the clear natural frequency peaks. The shape of these peaks can be used by either an interpreting specialist, or a knowledge-based computer program, to infer individual component dynamic characteristics such as damping and mode shapes.

Simultaneously with the impulse-correlated time-averaged sonic data accumulation, the structure-borne noise data due to normal operation of the machine is frequency-averaged to obtain operational response spectra. The ratio of pairs of the time- and frequency-averaged spectra, across the entire frequency range, provides new spectra which estimate the operational force spectra. Besides estimated the level of the operational forces, such as imbalance, the method allows the locations of peak forces observed to be obtained by triangulating from the locations of at least three sensors which are non-co-linear.

The approach can be implemented in systems of any level of sophistication, from being a manually applied field trouble-shooting technique, to being an information-gathering component of a self-contained artificially intelligent CBM system. For example, using such detailed information, an expert system can adaptively modify baseline dynamic models until the predicted and observed behaviors agree in all three spectra: modal time-averaged (exhibiting natural frequencies, mode shapes, and damping), structure borne noise and vibration which are not time-averaged (exhibiting net results which can be compared to expectations and pass/ fail criteria), and excitation forces (levels and locations, which can be compared to machinery performance models and test-stand measured norms). The adaptively modified models would allow very specific diagnosis of the source and nature of potential problems in the rotating machine. The expert system could then apply material, mechanical, and tribological models to perform prognostics.

A Detailed Description of the Proposed Approach: The first step involves the accumulation of vibration versus frequency spectra due to operational forces. For

this, it is necessary to obtain time-based data of vibration response to naturally occurring operational forces, and then perform a Fourier Transform on this data to convert it into vibration versus frequency, as is done in existing vibration-based CBM systems. Existing systems usually stop at this point, which is unfortunate since this information is necessary but not sufficient for reliable health management. Unlike existing systems, the new approach supplements the operational forced response data with response due to controlled and documented artificially imposed impulse forces.

The second step determines the machine's frequency response functions which relate to the individual components of the machine, and the attachment structures between these components. As part of this step, it is necessary to obtain time-based data of vibration response to artificial excitation by multiple impulses, where the beginning of the response "time window" from each impulse is triggered by the impulse. This is a primary innovation of the proposed new approach, and might be induced, for example, by a small electromagnetic collar encircling the shaft, or miniature piezoelectric-driven exciters. If actively controlled magnetic bearings are available, which is unusual but may become prevalent within the next decade, the artificial impulse might be induced by them without need of additional shaft exciters.

The purpose of the artificial impulse excitation is to measure, at several locations, the response of the machine's internal components to a known level of force over a broad frequency range. An impulse is ideal for this task because the Fourier Transform of an impulse results in relatively constant force versus frequency from very low frequencies up to a roll-off frequency. The briefer the impulse, the higher the frequency at which force "roll-off" occurs. However, there is a trade-off, in that the shorter the impulse is, the lower the level of force is at any given frequency. A shorter impulse could be made stronger to compensate for shorter duration, but this requires a larger exciter, implying significantly additional cost and weight, impractical for many applications. Large exciting force could also lead to excessive stress near the point of impulse application, particularly within shock-sensitive components such as mechanical seals and rolling element bearings. This means that the maximum force which is achievable in the impulse method has a practical limitation. In fact, unfortunately the practical force level is not high enough to directly compete successfully with the fluid system natural excitation operational forces in most machinery systems.

Therefore, a second innovation is proposed which involves signal processing performed to emphasize, within the total vibration signal, the system's response to the low level impulse versus its response to any forces that do not correlate with the timing of impulse, such as the naturally occurring forces. Prior to the Fourier transformation of the vibration response data sample taken during and for a fixed time following the artificial impulse excitation, the digitized vibration vs. time data signal is stored in a buffer, with the vibration level at each digital instant in time being stored at a different location in memory. The storage sequencing in memory is based upon the number of digital time steps since the initiation of the impulse. Following the next impulse data sampling, that sample is averaged with the previous sample, at each digital instant in time, where the memory location with which it is to be averaged (and at which the new average will then be stored) is defined on the basis of the number of digital time steps since the new impulse

had started. Statistically, this process leads to a zeroing of the memory buffers for any random vibration which does not correlate with the impulse, since such vibration has a much chance of being sampled as negative as its does of being sampled as positive at any given instant of time after the occurrence of the impulse. In this manner, a portion of the vibration at any given stored digital instant following impulsing will be reinforced (on average) if it is deterministically caused by the impulses, and will be de-emphasized (on average) if it is uncorrelated or incoherent with the impulses, such as would be true of vibration due to the naturally occurring operational forces.

The time-averaging process as defined above is repeated during additional impulses, with the sample averaging being performed cumulatively and linearly, in the sense that a new sample being averaged instant-by-instant into the buffer is weighted the same as each previous sample already in the buffer. In this manner, at the culmination of the averaging each sample's data receives equal emphasis (i.e. averaging is "linear"). The sampling and averaging is repeated enough times to be statistically significant (generally on order of 100 times for machinery systems while they are operating, based on the authors' field experience). At this point, Fourier transformation is performed, using an FFT analyzer. This process results statistically in the attenuation within the vibration frequency spectrum of response due to forces which do not correlate with the artificial impulse force (e.g. the natural excitation operational forces), and in the reinforcement of response due to the artificial force. After a sufficient number of impulses, the resulting time-averaged vibration spectrum divided by the artificial impulse force spectrum (based on the exciter force, which is tracked and stored separately from the vibration response) provides a good approximation of the true frequency response function or "FRF" of the system, as referenced to the locations where the impulse loading was performed and where the response to it was sensed. The FRF is basically the vibration response at one location per unit excitation force at another

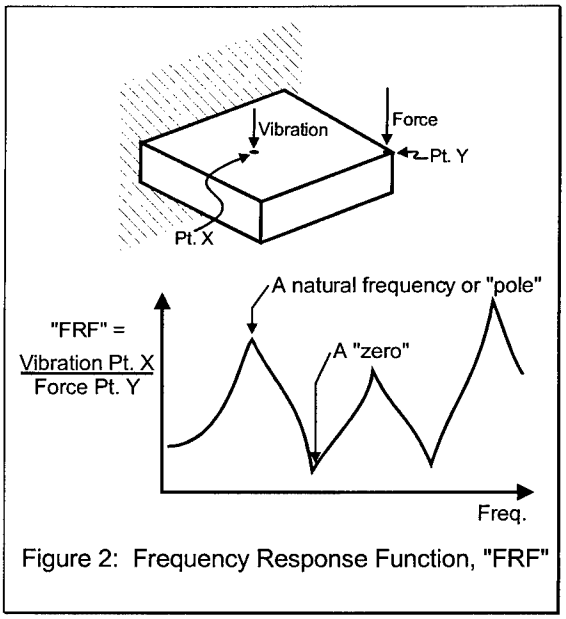


Figure 2: Frequency Response Function, "FRF"

(or the same) location, as a function of frequency. A descriptive example of an FRF plot, or "spectrum", is given in Figure 2.

The peaks or "poles" (i.e. the natural frequencies) of this FRF vs. frequency spectrum, the valleys or "zeros" (i.e. the frequencies at which vibration is a minimum), and the shapes of the peaks and valleys that form the individual poles and zeros, are the result of the stiffness, damping, inertia, and connectivity of the machine's subsystems, components, and mechanical supports and connections. On an individual basis, these in turn depend upon whether or not, for example, a crack has developed in an area of high stress, or a pump seal clearance is a certain value, or a bearing preload is being maintained. As components such as these wear, fatigue, embrittle, or loosen, the "map" of FRF vs. frequency pole/zero positions and shapes will change in a generally predictable, if fairly complicated, manner. Therefore, in principle it is possible to "read this map" to determine the mechanical health of the machine, its individual subsystems, and the individual components of the machine and system.

Besides FRF's, other "maps" are available for interpretation relative to mechanical changes and health implications once an accurate FRF is determined. For example, the "lag" between when the impulse force is applied and when time-averaged response at a given frequency occurs is called "phase". Along with FRF amplitude, phase is a function of frequency, and this function will change shape as mechanical deterioration occurs in a rotating machine. This can be tracked directly, or through looking for pattern shifts in graphs more sensitive to mechanical change, such as root locus plots of the vibration viewed as a complex number, with the imaginary versus real components plotted against each other.

In principle, the impulse modal test could be performed in less time and perhaps without the rigor of time-averaging if it were done with the system not running. Systems already exist which can do this as part of a trouble-shooting procedure, if desired. However, these existing modal testing systems cannot be used in-service or on the test stand, and being able to do so is a significant advantage of the approach discussed in this paper. Furthermore, the FRF of the rotor system obtained on a non-operating machine generally has insufficient accuracy when the system is not operating. For example, casing wall cracks may be tightly closed, and the rotor dynamic coefficients of the bearings, seals, and impellers or bladed disks are much different in a non-operating system versus one during operation. Finally, by obtaining precise FRF's from several locations during operation, the new approach is able to determine with reasonable accuracy the operational forces at key frequencies within the various machinery components and subsystems, throughout the machine's operating range, or on the test stand, as described in the next step. No other system is able to provide this information, unless massive amounts of additional load cell and pressure instrumentation is implemented.

The third major step in the new approach is to divide the frequency spectrum due to operational forces by the FRF due to the impulses, to obtain an estimate of the operational force vs. frequency spectrum, referenced to the several instrumented locations. It is then possible to perform tomographic triangulation to locate the sources of significant forces over selected frequency ranges, and re-reference the force vs. frequency spectra to those locations.

Before the FRF is interpreted, however, the new approach returns to the vibration caused by the machine's operational forces, rather than by the impulses. This operational force vibration is the only signal considered by existing vibration-based CBM systems, in either aerospace or industrial applications. In such existing systems the vibration versus frequency spectrum is either directly provided to an operator or maintenance technician for interpretation, or is reviewed by a knowledge-based computer program to determine if overall vibration level or vibration at a specific frequency is higher than some pre-programmed criterion. In computerized systems, the occurrence of above-normal vibration at specific frequencies may be interpreted by a subroutine, which then lists what specific mechanical or operational problems might exhibit most of their energy at the noted frequency. While useful, such a list of potential problems generally includes too many potential offenders to allow specific diagnosis, and therefore is unreliable about whether or not the vibration indicates that the system is truly in danger. Because of the non-uniqueness and "fuzziness" of its interpretations, such a system seldom is able to reliably flag a specific offending component, or to be precise about what is wrong with any component that it is able to flag. Therefore it cannot make credible projections of remaining life.

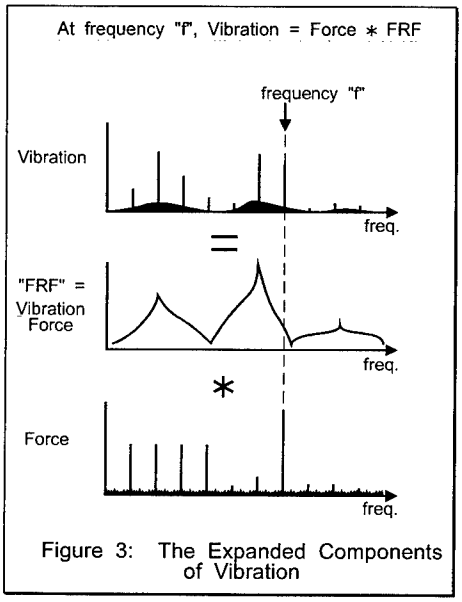


Figure 3: The Expanded Components of Vibration

The new approach also considers the operational force spectrum, but is able to be much more quantitative in its interpretation of it by using the FRF spectrum generated by the impulses. The natural excitation response spectrum is divided (on a frequency-by-frequency basis across the digitized frequency spectrum) by the FRF referenced to the same measurement location, thereby estimating the operational excitation force spectrum cross-referenced to the location of the impulse excitation. By using small exciters in at least three locations which are not co-linear or co-planar, triangulation may be used by a tomography process to determine the specific origin of the various forces making up the operational force spectrum. This is true of both relatively narrow-band forces such as pump or steam turbine blade passing frequency, as well as broad-band forces such as gas turbine combustor pressure pulsations or compressor rotating stall cell frequencies.

The fourth step in the new approach is to compare force and FRF test spectra to analysis predictions, and perform analysis iterations, guided by either a

knowledgeable specialist, or in principle by knowledge-based computer software, until they agree. With such an approach as outlined above, the original operational vibration versus frequency spectrum can now be expanded into two more fundamental components as shown in Figure 3, i.e. the amount that the machine responds to a given level of force at a given frequency, and the amount of force that is available to in a given subsystem component to excite the system at that frequency. This expanded information is able to be directly interpreted by the specialist or knowledge-based computer program.

The fifth step in the new approach is to identify specific faults in subsystems and components based on changes required for the analytical models to bring them into agreement with the test data. The force levels at various frequencies can be evaluated based upon their probable effects on system stresses and utilization of the machine's running clearances. The peaks in the FRF spectrum represent system natural frequencies, and the shape of these peaks provides quantitative information concerning modal stiffness, damping, inertia, and integrity of the connectivity of the overall machine, and its individual subsystems and components, as discussed above.

The final step is to determine the rate of deterioration and remaining useful component life, by implementing component tribological models. Models based on linear elastic fracture mechanics, bearing wear rate versus existing clearance and load, and similar situation-specific and component-specific analyses, can be used to predict component life based on deterioration deduced by a specialist or a knowledge-based computer program from test/finite element comparisons.

Utilization of the New Approach: The overall interpretation of machinery health now can be reliably specific concerning its conclusions. For example, for turbine driven boiler feed pumps, what is the level of imbalance at the pump-side and turbine-side? Is the level within expected limits? At the indirectly measured imbalance level, what is the rate of wear in the bearings? What is the level of associated rotor vibration? Added to vibration from other sources, will this cause rubbing at the pump's annular seals? If so, how much clearance loss is likely? How much have the stiffness and damping matrix values of the individual bearings and seals already changed versus design, based on the shift in associated poles and zeros on the FRF plots? What do tribological (i.e. lubrication and wear) models indicate is the current deterioration rate and extrapolated remaining life of these bearings and seals?

On a wide variety of industrial field consulting projects, the authors have successfully performed the type of experimental data vs. analytical model comparisons described above, in order to locate and understand problems on industrial machines and fluid systems. In each of those cases a human specialist (either the authors or one of their associates) was present to guide the analytical iterations based on detailed knowledge of the machines and processes, combined with pattern recognition insights. In most applications, generally time and funding is not available for such direct use of human expertise, or even significant human interactivity with the computer program. Fortunately, the proposed approach outlined in this paper can be fully automated, with control-room manual over-ride. In addition, pre-programmed patterns and "fuzzy logic" pattern

recognition principles can be used to minimize the computing time required in order to reach a reasonable judgment concerning the reason for the changes.

CLOSURE: There have been significant strides in condition-based maintenance systems over the last two decades, but they still are not reliable enough to be useful without substantial human interfacing and intervention. The sonar-like approach discussed in this paper has shown its ability to obtain unique, accurate, and reliable information concerning the health of various machines, and can do this continuously while they are operating. The approach has succeeded in determining the cause of over 150 difficult machinery vibration problems, and in providing reliable guidance concerning remaining time-to-failure and problem solution. The new method should be further explored, and its ability to be applied on an automated basis should be investigated. In the meantime, its usefulness as a field troubleshooting tool will continue to be useful, particularly in critical applications or in the solution of problems which are chronic and difficult to diagnose.

PROGNOSTIC ISSUES FOR ROTORCRAFT HEALTH AND USAGE MONITORING SYSTEMS

Carl S. Byington, Susan E. George, and G. William Nickerson

The Pennsylvania State University
Applied Research Laboratory
Condition-Based Maintenance Department
State College, PA 16804

Abstract: A description of health and usage monitoring systems (HUMS) is provided along with a brief justification of the need for transition to condition-based maintenance (CBM) for safety and affordability in rotorcraft operations. A set of issues is posed for mechanical diagnostics and prognostics intelligence in the evolution of HUMS and transition to CBM. A notional scenario outlining requirements for detection, diagnostics, and prognostic system design is presented. A methodology for setting acceptable false alarm rates (both lower and upper bounds) is proposed along with a case for a human-system interface for shared decision making in the ultimate HUMS.

Key Words: Condition-Based Maintenance; Diagnostics; Health and Usage Monitoring Systems; Human Systems Interface; HUMS; Prognostics.

Introduction: Presently, the amount of maintenance on flight critical aircraft components is excessive while the amount done "on-condition" is minimal. Safety-of-flight considerations have demanded that critical components be maintained "before they can fail", the definition of a *preventive maintenance* philosophy. This approach has resulted in high maintenance costs as expensive components are removed and discarded long before their useful life is actually consumed. Unfortunately, despite the conservative nature of component safe life estimates, failures continue to occur at an undesirably high frequency. Implementation of advanced rotorcraft Health and Usage Monitoring Systems (HUMS) will facilitate transition to a Condition-Based Maintenance (CBM) philosophy with a resultant decrease in maintenance costs for rotorcraft. But, such a transition will occur when (and only when) accurate detection, diagnostic, and *especially* prognostic capabilities become a reliable part of the ultimate HUMS.

The Condition-Based Maintenance philosophy stipulates maintenance of equipment only when there is objective evidence of an impending failure on the particular piece of equipment [1]. Rotorcraft HUMS can enable CBM in that they monitor the condition and observables of faults in critical mechanical components and systems by detecting abnormalities in sensor data; in effect acting as a vigilant "watchdog" for evolving faults. Detection of faults in the drivetrain, engine, or rotor system is accomplished via vibration monitoring, oil debris monitoring, and exceedance monitoring of temperatures, pressures, shaft speed, torque, and strain. Usage monitoring estimates the consumed life of dynamic components based on specific flight regimes and environmental conditions to which the helicopter has been subjected. The combination of fault detection and diagnosis coupled with reliable usage monitoring can provide the basis for a reliable prognostic capability enabling CBM.

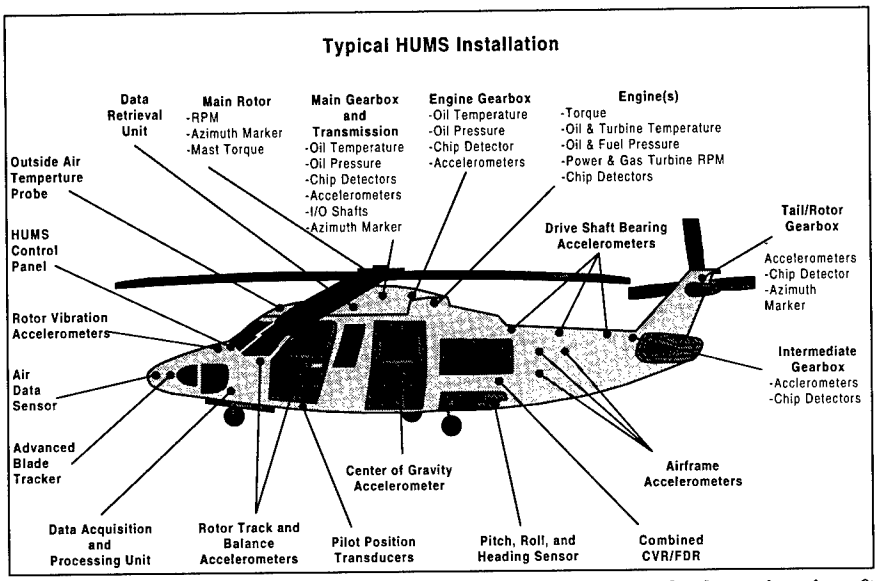


Figure 1. A typical HUMS installation is complex and affects nearly the entire aircraft [adapted from the Teledyne Controls/Stewart Hughes poster "The World's First Certified Helicopter Health and Usage Monitoring System for Bell 412"]

HUMS acquires its data from sensors mounted throughout the aircraft and from existing flight instrumentation, Figure 1. Diagnostic, usage, and advisory information concerning critical flight components can be presented by HUMS to the pilot and crew through a human computer interface. HUMS can also track maintenance actions and provide post-flight diagnostic capabilities through the processing of flight data at the ground station. The significance of augmenting existing HUMS in-flight diagnostic capabilities will be discussed. For further background HUMS information, consult References [2] and [3] .

HUMS Motivations: The primary motivations for the development of rotorcraft Health and

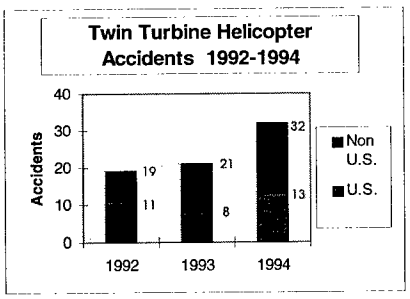


Figure 2. Twin turbine helicopter accidents have increased dramatically over the past few years [2].

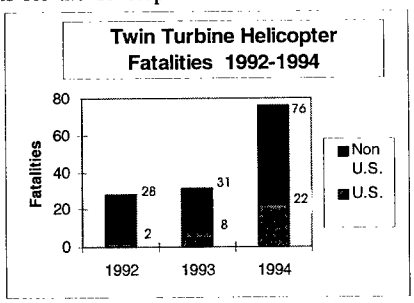


Figure 3. There are more fatalities resulting from twin turbine helicopter accidents than in previous years, with nearly one third occurring in the U.S.[2]

Usage Monitoring Systems are increased safety and reduced life cycle costs. Helicopters are subject to extreme dynamic loads and vibrations that impact their engines, drivetrain, airframe, avionics, and rotors. In addition to the mechanical faults typical of any performance-driven application with little design margin, the highly-variable and extreme loads that are generated during normal rotorcraft operation tend to cause failures that propagate rapidly and are nearly impossible to predict. Unanticipated, in-flight failure causes helicopter accidents, resulting in a fatal accident rate per mile flown an order of magnitude higher than that for fixed wing aircraft [2]. In the United States, the annual number of rotorcraft accidents is over 2.5 times greater than the number of fixed wing aircraft accidents despite dramatically fewer helicopter flight hours vis-à-vis fixed wing.

Worldwide, an alarming increase in twin turbine helicopters accidents and fatalities has be seen This increase is illustrated in Figure 2 and Figure 3 [2]. In 1994, 76 people died in twin turbine helicopter accidents with nearly one-third of them in the United States.

Implementation of HUMS can improve both civilian and military rotorcraft safety but the military need may be more urgent. HUMS increases the safety level of the rotorcraft by continually monitoring the condition of critical components - acting as a "watchdog". Military rotorcraft are often subject to more extreme flight conditions resulting in increased likelihood of premature mechanical failure. Chamberlain [4] illustrates the potential impact of HUMS on United States Navy (USN) and Marine Corps. (USMC) rotorcraft safety by citing historical data from the Naval Safety Center on Class A mishaps for rotorcraft. From 1980 to 1990, a total of 204 helicopters were lost with 57 fatalities. Of the 204 losses, 85 were deemed to be mechanical systems related. Data from 1992 to 1993 indicates 29 mishaps with 14 attributable to engine, drivetrain, or rotor system failure. A distribution of these mechanical faults is provided in Figure 4. A study reported by Chamberlain evaluating these mechanical faults determined that 64% (1980-1990) and 79% (1992-1993) of them might have been prevented with current HUMS technology (Refer to Figure 5). As the U.S. Military's rotorcraft fleet ages, more unpredictable failures are likely to occur. If these rotorcraft are retrofitted with current HUMS, some failures would be detected in time to prevent many future military accidents and fatalities.

Although HUMS is a relatively new technology, there have already been numerous instances where HUMS has detected a fault and averted a potential tragedy. In March 1993, the world received proof of HUMS effectiveness when a Stewart Hughes Ltd. HUMS detected a serious problem on a Norwegian Boeing 234 Commercial Chinook and resulted in the first HUMS cancellation of a commercial flight [5]. Investigation found that a 10 mm bolt on the number two cross-shaft coupling between the engine and a gearbox had broken. It is not knowable whether or not this bolt would have caused an accident if the flight had proceeded without the warnings of the health monitoring system, but the sheared bolt had the potential to cause a lethal accident. In addition, another operator using the Teledyne/Stewart Hughes HUMS on about 20 rotorcraft has reported success [2]. Since installation in that operator's rotorcraft fleet, twenty faults were detected and subsequent failures prevented. Four of the faults were considered potentially catastrophic and six would have caused costly maintenance. Other HUMS

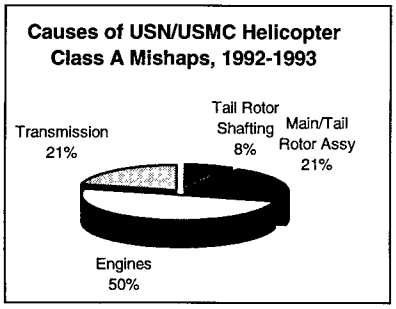


Figure 4. Failures in various helicopter mechanical systems caused over 40% of Class A Mishaps [4].

successes have since been recorded [2][6][7][8].

HUMS Added Capabilities: HUMS improves safety through means other than fault detection. HUMS systems can reduce pilot workload by automating some routine pilot tasks and allowing them to concentrate on other important aspects of the flight. In addition, if a HUMS system is permitted a certain level of decision-making, it can correct some human errors and possibly prevent an accident from occurring [8]. In some cases, a health monitoring system can detect problems and respond to them before the pilot becomes aware of their existence. Depending on the type of problem detected, HUMS can perform corrective actions needed to rectify the situation much faster than a pilot. For example, if a failure occurred in the primary hydraulic or electrical system or in an engine, HUMS could assure a quick switch to the correct secondary systems [8]. Since there are numerous instances of pilots shutting down the wrong engine in the event of an engine fault, for example, owing to ambiguous and misleading indications, this could make the difference between life and death for the rotorcraft's passengers and crew.

There has been much deliberation over the allocation of responsibility for decision-making in the aircraft [9]. How much control over decision-making should be assigned to the HUMS computer system? Certainly, some situations might be better handled by a computer while others would be resolved more effectively by the pilot. Similar issues had to be addressed for fixed wing aircraft with the introduction of unstable aircraft such as the X-29. This aircraft is capable of flying only with a computer making continual in-flight adjustments to the pilot's actions - "fly-by-wire". Without the computer's assistance, pilots can only fly the aircraft for a few seconds before it becomes unstable and completely uncontrollable. It is certain that fly-by-wire was not adopted by the pilots because it was "new technology". They accepted it because it met their immediate operational requirements.

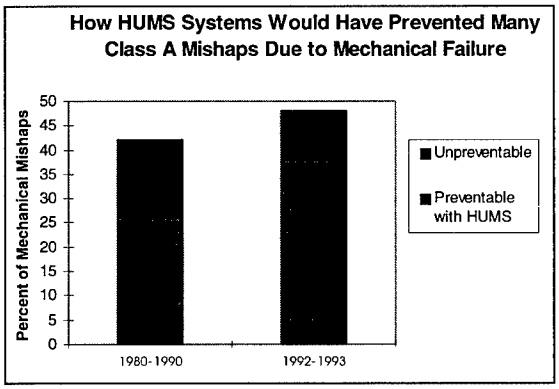


Figure 5. If current HUMS technology had been available during the past 15 years, many Class A mishaps caused by mechanical failure could have been prevented [4].

The best decision-making would result if the observations and recommendations of both the pilot and the HUMS system could be integrated before determining the seriousness of a fault condition or the course of action in response to that condition. The helicopter human system interface must be designed to effectively relay HUMS information to the aircrew, to permit the pilot to query HUMS for more details following an alert, and to allow the aircrew to combine their knowledge of the situation with HUMS information. Issues associated with achieving this objective are the focus of this paper.

Benefits and Costs: Although the value of HUMS has not been fully quantified, the benefits extend beyond improved safety and have generally been accepted by manufacturers and operators. Operating costs for helicopters are extremely high. Twenty-nine percent of the operational cost can be attributed to insurance and twenty-four percent to maintenance [2] as shown in Figure 6. Operating costs would be significantly reduced if there were a way to

decrease insurance premiums or the amount of maintenance required. Current insurance costs are largely due to high liability awards for accidents and high levels of real or perceived risk associated with helicopter flight. As HUMS matures and prevents mishaps, confidence will increase and resulting decreased risk will be reflected in lower insurance premiums.

Excessive maintenance expenditures result from frequent maintenance checks and mandatory part removal after a specified operational life has passed. Continuous monitoring of critical components means more impending failures are likely to be detected. Because critical components are monitored between the required maintenance checks, it should be possible to safely extend the interval between costly maintenance overhauls. In addition, components with *remaining useful life* would not be replaced simply because their statistically determined life limit has expired; instead, they could remain in service for a reasonably extended period of time or until the health monitoring system detects a problem. Extension of the duration between maintenance checks would also result in greater aircraft availability, particularly benefiting helicopter fleet owners.

Expected life cycle costs of HUMS arise from the initial purchase, installation and subsequent operation of the system. Multiple sensors, necessary wiring, an onboard data processor, and a ground station computer (GSC) comprise a list of the major HUMS components. Operational costs include the manpower needed for the ground crew to collect and interpret and archive the data as well as the general maintenance required for HUMS readiness. Because helicopter models and designs differ, HUMS systems need to be customized for each type of helicopter, making installation more expensive [2].

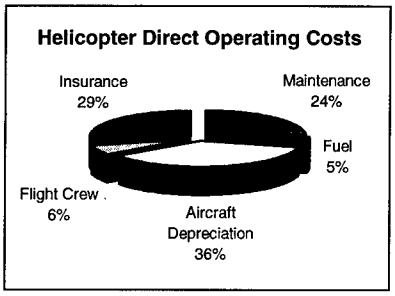


Figure 6. Maintenance and Insurance expenses contribute significantly to a Helicopter's Direct Operating Costs [2]

The best opportunity to install HUMS arises when the helicopter is already grounded for a maintenance overhaul. This will reduce the cost of installation. In older rotorcraft, especially those with only a short life remaining, the cost of installation and operation may not be justifiable but it should be analyzed for the particular case. In many newer rotorcraft manufactured for civilian use, a HUMS system is installed during assembly.

Civil aviation authorities around the world are increasingly requiring accident data recording (ADR) systems which usually consist of flight data recorders (FDR) and cockpit voice recorders (CVR). Now some manufacturers are including a HUMS system in the package along with more advanced signal acquisition and processing functions [10] . By installing the ADR system and HUMS simultaneously, implementation costs are reduced. HUMS is surely a welcome addition to the accident investigation team because its data record can assist in determining the causes of accidents. Unlike civil aircraft, military aircraft are generally exempt from ADR requirements, and military organizations may choose to install HUMS without ADR due to acquisition budget constraints.

Diagnostics and Prognostics Intelligence: The real power and the opportunity to transition to the safer and more cost-effective Condition-Based Maintenance philosophy for rotorcraft will only be realized when HUMS systems are intelligent and capable of prognostics. Future HUMS systems will require software intelligence for the detection of faults, failure mode diagnosis, and prognostic projection. The machine intelligence must possess the ability to extract results and

convey information to the operator. Figure 7 illustrates the levels of processing that describe the prognostic CBM process.

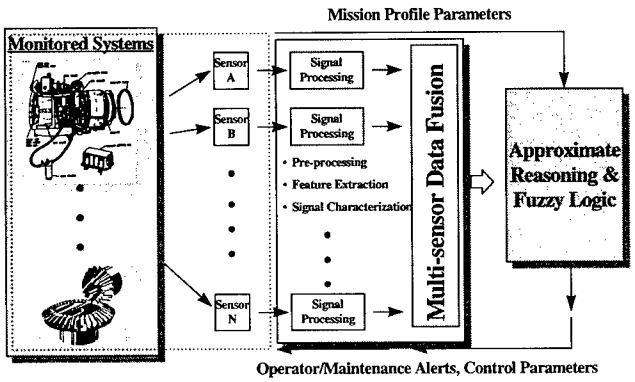


Figure 7. Data from monitored systems is processed at several levels and fused together to provide a better picture of a machine's condition.

CBM data processing includes sensor data association, sensor validation, state vector estimation, and anomaly identification [11]. Hall [12] has described many multi-sensor data fusion methods that are directly applicable to the CBM problem. Research in this area continues with the ONR-funded Multi-Disciplinary University Research Initiative in Integrated Predictive Diagnostics and is being applied to projects ranging from single gear meshes to gas turbine engines. Methods to extend the feature vector and the observable operating range are being explored. If these efforts are successful, the promise of false alarm reduction and increased measurement confidence through sensor confirmation and validation may be reaped.

While corroboration through data fusion provides robustness and the synergy associated with the combination of multiple sensors, the essence of fault diagnosis is based on abnormal pattern recognition. Data deviating from some defined normal pattern usually indicates the possible presence of a fault. Traditionally, statistical (similar to process control) methods such as regression analysis have been used in fault diagnosis to model and analyze non-linear dynamic systems. This method has demonstrated its limits when it comes to detecting slowly evolving faults and abrupt ones. Hence, the application of fuzzy logic and artificial neural networks (ANN) are being developed to improve HUMS pattern recognition and fault diagnosis. But, the key issue is to define the functional requirements that HUMS must provide to be a useful operational tool for the transition to CBM.

Design for Prognostics: To satisfy the operators' requirements, a target for future HUMS is proposed as:

> *The ability to detect all flight critical faults at least 2 minutes prior to their reaching operational limit and 99% at least 10 minutes before operational limit and 90% a minimum of two flight hours prior.*

These goals imply that 90% of all failures will be caught and repaired before the aircraft leaves the terminal. Of the 10% that are detected in flight, 90% of those are detected with a minimum of 10 minutes to respond. With the remaining 1%, a minimum of 2 minutes will be provided. These are the so-called "3-sigma" failures that progress extremely rapidly and provide no detectable indication prior to their ultimate failure. These could be faults induced by foreign object or battle damage.

It is believed that ongoing research will be able to increase the probability of detection prior to departure and to increase the effective warning period for faults detected in flight, but the above goal represents optimistic projections of future performance. It must be remembered in

considering this performance that flight critical failures occur in helicopters on the order of 1/100,000 flight hours, so we are addressing a rare event which makes quantitative performance measures difficult.

All outputs of the HUMS must be bounded with certainty or confidence limits to be useful to the human and to downstream analytical processes. These measures will necessarily be calculated with the detection, diagnostic, and prognostic algorithms and be continuously updated.

The following notional scenario describes the manner in which an "Ultimate HUMS" might determine and relay information about the evolution of a fault. Recognize that the scenario presented is entirely notional and the capability it represents may never be achieved. But, it's purpose is to articulate a view of a desirable future state and promote discussion to define the most operationally useful system possible.

FAULT DETECTED: [*Italicized* text represents the HUMS output.]

The first step in the reasoning process is to detect that a fault has initiated and become observable. This condition represents a variation from "normal" and will trigger a set of more detailed analyses to establish a confidence and diagnose the nature of the fault. Thus, the detection call might be made as follows:

"Drivetrain fault detected, 99% certainty of fault."

FAULT DIAGNOSED:

Diagnosis differs from detection in that diagnosis implies that the specific fault detected has been identified and quantified with respect to its severity. The ultimate HUMS should be able to perform this diagnosis with a high degree of certainty at least to the line-replaceable unit (LRU) level. It may be important to the prognosis function that the diagnosis be more specific in order to predict the remaining useful life, but the operator is most concerned with having that knowledge available down to the lowest LRU for maintenance planning. The ultimate HUMS should be able to achieve this diagnosis function with 95% accuracy (counting false alarms and misses) for the faults that are currently detected on the ground. At the LRU level, the diagnosis should be presented concurrently with the fault detection message illustrated above. In order to refine the diagnosis below the LRU level, additional time may be required for further analysis by the system.

Again, the reliability of the diagnosis will need to be calibrated with certainty measures. Thus, the detection call would be expanded as follows (new content shown in ***bold italics***):

"Drivetrain fault detected, 99% certainty of fault. ***Diagnosis: main rotor transmission planet gear spalling, 2% (+/- 0.5%) of tooth area, 90% certainty.***"

EVOLUTION PREDICTED (Prognosis):

Prognosis is the goal of the Ultimate HUMS system. The term means: reliable and accurate prediction of the remaining useful life of a component in service. Thus, true prognosis requires not only a prediction of the time to failure but also a measure of the certainty and error bounds on the projection. With this information, the aircrew can have (but may choose not to use) the opportunity to modify operations to extend component life. The worst case projection of remaining useful life must be set at a high level of confidence, probably 99.5% based on three standard deviations (3σ) below the mean prediction. This capability does not now exist for even relatively simple devices, let alone complex helicopter gearboxes. The true power of HUMS and mechanical diagnostics to impact operational safety and cost lies in a realized prognostic capability.

The minimum reliability (certainty) in the projection will be established prior to flight, i.e., for a given fault, the system will not present a prognosis without achieving the prescribed level of certainty. It is extremely important not to present the aircrew with a projection that is longer than will be realized under worst case conditions. It is also important, however, as we move to prognosis, that we not project too conservatively lest we continue to "throw away" useful life of components and increase operating costs.

The diagnosis call from above would be expanded as follows with prognostic information:

"Drivetrain fault detected, 99% certainty of fault. Diagnosis: main rotor transmission planet gear spalling, 2% (+/- 0.5%) of tooth area, 90% certainty. ***Remaining safe flight 10 to 40 minutes.***"

Setting Acceptable False Alarm Rates: False alarms are the nemesis of all mechanical diagnostic systems. On the one hand, high sensitivity to symptoms of an evolving fault is desired to give the earliest possible warning. On the other hand, high sensitivity can lead directly to high false alarms. The threshold and sensitivity settings are therefore trade-offs that must be carefully balanced between the consequences of a failure and the implications of a false alarm, both operationally and on the confidence the user has in the system. From an operational perspective, statistical decision theory will most likely be applied to account for the consequences of failure versus the consequences of false alarms in light of their respective frequencies. In helicopter operations, there are many situations where an extremely low false alarm rate is imperative because the consequences of a false alarm induced immediate landing (statistically) exceed the consequences of a failure in flight. In that case, monitoring (or certainly diagnostics) is not advisable unless an acceptably small false alarm rate is achievable. The statistical decision theory analysis will provide the lower bound on the acceptable false alarm rate as a function of the monitored system.

The upper bound on acceptable false alarm rate is established by the user. Even if analysis shows that being wrong frequently (say 40%) is acceptable from a statistical decision theory cost/benefit perspective, there is a threshold above which the user will no longer have any confidence in the system. There is some research indicating that the psychological limit is in the vicinity of a 15% false alarm rate, i.e., users will believe an alert until the rate at which it proves false exceeds about 15% [13].

These two bounds establish proposed design limits for false alarm rate in the HUMS. The lower level of acceptable false alarm established by a statistical decision theory analysis of the system. The highest level of acceptable false alarm at 15%, regardless of the results of the system analysis. For the purposes of this paper, it must be assumed that the HUMS system will achieve the required minimum acceptable false alarm rate through ongoing development perhaps by including input from the aircrew.

Human-System Interface for Shared Decision Making: Fault detection, diagnosis, and prognosis information collected in flight must be presented to the maintainer on the ground and should be available to the aircrew. Most current HUMS system have a fairly efficient means to download the information from the aircraft and interface it to the maintenance infrastructure for scheduling and parts ordering. On the other hand, current HUMS have extremely limited interfaces to the aircrews although the information to be communicated could be life-saving. Future HUMS will necessitate a much richer interface applying most appropriate technology to inform the aircrew and to receive additional information from them. Due consideration of context and accuracy are important because when a fault has been detected and the crew is notified (in general), the information will likely cause an emergency response. Thus, it is important that the information presented be unambiguous and easily understood and assimilated by the aircrew.

One of the main differences, in terms of an evolution towards future HUMS, is that the information flow will no longer be uni-directional from the system to the cockpit. It will necessarily be bi-directional with the aircrew having the ability to query the HUMS to gain additional corroborating information or to provide information to the HUMS system that can only (reasonably) be provided by the human. For example pilot-identifiable changes in the performance of the aircraft like a "sluggish stick" could be useful to, but not directly observable by, the HUMS system. With such a level of interaction, it is likely that additional methods of communicating the diagnostic and prognostic information using tactile as well as visual and aural senses will be available. The information portrayed in the detection, diagnosis, prognosis scenario above probably would be most effectively presented visually using some sort of advanced graphical interface.

A key HUMS design issue that must be addressed in the human system interface development and especially in the transition to onboard HUMS/aircrew interaction is: How much corroborative information the aircrew should be able to (wants to be able to) access in flight? In principle, it would be desirable to provide any information the aircrew wants. In practice, this will not be possible or practical. The information that HUMS will use to detect, diagnose, and predict failure evolution will be extremely technical in nature and will have been subjected to a high degree of automated analysis by the time that an alert is posted. Thus, the type of corroborative information that the aircrew wants must be defined in advance in order to make it available.

A probable requirement is that the HUMS system will be required to provide information to the human system interface regarding the basis for its conclusion. HUMS must, for example, be able to identify that it predicts 10 hours to failure of the planet gear in the main rotor transmission due to spalling because:

- HUMS has detected debris in the sump consistent in quantity and morphology with gear spalling.
- HUMS has detected a vibration signature that is indicative of a surface flaw on the planet gear.
- HUMS analysis of the amplitude and spectral characteristics of the vibration signature and rate of debris generation indicates that the severity of the spalling is 2% of the tooth area.
- The spall limit for that gear (based on design and experience) is 4% of tooth area.
- Under worst case conditions spalling has been seen to progress at 0.2% of area per flight hour.

This information would, of course, be provided from the HUMS system to the interface in a condensed format computer message using a pre-determined message format. The objective of the human-system interface should be to decode that message into "human terms" and maintain the logic and processes to allow the aircrew to ask "why?" of the HUMS system. This would likely be a hierarchical inquiry starting with the fault alert by the on-board diagnostic system and progressing to the most detailed information about the nature, severity, and context of the fault. The aircrew would ask repeated questions until they have achieved a satisfactory level of comfort with the HUMS prognosis. Thus, the integrated HUMS interface must provide the means to place the results of these analyses in context and in a format that the aircrew is able to digest in light of training, experience, and workload.

The real promise of the on-board diagnostic system will come when the system and the aircrew can act in concert to provide operational alternatives and their impact on remaining safe flight

time. For example, in the situation described for spalling of a planet gear in the main rotor transmission, reducing power might dramatically (orders of magnitude assuming a logarithmic S-N curve) increase the time to failure. Future HUMS could be able to accept input from the aircrew to assist in determining alternative remaining life scenarios. In general, anything that the aircrew can provide regarding future operations will allow the HUMS to extend its prediction of remaining useful life since its basic prediction is based on a "worst case" expectation, i.e., that the aircraft is flown in its most severe loading regimes for the part in question. This balancing act will be an area where close cooperation between the human-system interface and HUMS designers will be required to produce an integrated on-board diagnostic/prognostic system. The design (and hence the cost) of the system will be heavily driven by the level of this type of processing it is required to perform.

Summary: A review of Health and Usage Monitoring Systems and a description of HUMS functions were presented. The benefits of HUMS systems are well documented in both civil and military communities. In addition, this paper discussed issues which are expected to arise as prognostic capabilities and on-board interfaces are incorporated into HUMS. It is hoped that the paper will be useful in defining a future HUMS which will expand current diagnostic and prognostic capabilities and more effectively provide this information to ground maintenance personnel and aircrews. The means by which this can be achieved is by active and continuous discussion of the propositions offered.

Acknowledgment: This work was supported by the *Multidisciplinary University Research Initiative for Integrated Predictive Diagnostics* (Contract Number N00014-95-1-0461), the grant *Advanced Human-System Interface to Integrated Diagnostics in a High Stress Environment*, (Grant Number N00014-95-1-0941), and a Phase I SBIR Award to CHI Systems, Inc. *HUMS Interface System (HINTS).* All are sponsored by the Office of Naval Research.

References:

[1] C. S. Byington and G. W. Nickerson, "Technology Issues for Condition-Based Maintenance", *7th Annual Predictive Maintenance Technology Conference,* 1995.
[2] J. Land and C. Weitzman, "How HUMS Systems Have the Potential of Significantly Reducing the Direct Operating Cost for Modern Helicopters Through Monitoring", *Teledyne Document, presented at the American Helicopter Society 51st Annual Forum and Technology Display*, Ft. Worth, TX, May 1995.
[3] P.W. Stevens, D. L. Hall, and E. C. Smith, "A Multidisciplinary Research Approach to Rotorcraft Health and Usage Monitoring", presented at the *52nd Annual Forum of the American Helicopter Society*, 1996.
[4] M. Chamberlain, "U. S. Navy Pursues Air Vehicle Diagnostics Research", *Vertiflite*, March/April 1994.
[5] G. Marsh, "The Future of HUMS", *Avionics Magazine*, p. 22-27, February 1996.
[6] D. Parry, "Evaluating IHUMS", *Avionics Magazine*, p. 28-32,February 1996.
[7] "Health and Monitoring System Enhances Helicopter Safety", *Aviation Week and Space Technology* (New York), v 139, n 12, p. 94, September 20, 1993.
[8] D. G. Astridge, "HUM-Health and Usage Monitoring of Helicopter Mechanical Systems", *British Journal of Non-Destructive Testing*, v 30, n 6, p. 403-408, November 1988.
[9] "Human-Computer Interface to an Integrated Diagnostic System in a High-Stress Environment", Workshop Results, August 3-4, 1995.
[10] K. F. Fraser, "An Overview of Health and Usage Monitoring Systems (HUMS) for Military Helicopters", *Australian Defence Science and Technology Organization, Report No. DSTO-TR-0061*, 1994.
[11] M. E. Kotanchek and D. L. Hall, "CBM/IPD Process Model", *ARL TM 95-113*, November 1995.
[12] D.L. Hall, "Applications of the JDL Data Fusion Model to Condition-Based Maintenance Systems", *DFS-94 Tri-Service Data Fusion Proceedings*, October 1994.
[13] Murray Gell-Mann, *The Quark and the Jaguar*, pp 258 - 259, 1994.

HIERARCHICAL FAILURE SIMULATION FOR MACHINERY PROGNOSTICS

George A. Lesieutre
Penn State University
233 Hammond
University Park, PA 16802

Lei Fang
Penn State University
233 Hammond
University Park, PA 16802

Usik Lee
Inha University
253 Yonghyun-Dong, Nam-Ku
Inchon 402-751, South Korea

Abstract: A "prognostics" capability is an essential element of a condition-based maintenance scheme. Prognostics involves *monitoring* to ascertain the current health of a machine in combination with *predictive modeling*, and implies the ability to confidently predict remaining life under a range of operating conditions so as to guarantee operation for a minimum time under extreme conditions. Development of efficient numerical models of material damage accumulation and component failure are key parts of a prognostics capability. Such models need not approach perfect fidelity, as errors can be detected by monitoring, but must be accurate enough to guarantee safe operation over some time window into the future. A hierarchical modeling approach is being developed to address this need, with the following features : (1) local material failure is modeled approximately using continuum damage mechanics; (2) component failure is modeled through numerical solution of representative boundary value problems; and (3) system failure is modeled dynamically by representing individual components as discrete elements having associated damage states. Damage mechanics solutions for individual components are used to define characteristic "component damage modes," which are used to construct low-order component models for use in system simulation. These simulations are used to identify system failure modes and accompanying observables, as well as to assess remaining life.

Key Words: Condition based maintenance; condition monitoring; prognostics; damage mechanics; damage evolution; damage modes; remaining life.

INTRODUCTION: Machinery prognostics involves a combination of monitoring and prediction, with the goal of confidently and accurately predicting remaining life. If safe operation for some minimum time under extreme conditions cannot be guaranteed, maintenance is scheduled. Machinery prognostics is currently of great interest in the machinery monitoring community [1-4]. The accurate prediction of machinery failure can be thought of as being broadly founded on two approaches:

(1) History-based. If a manufacturer (or an operator) of a particular kind of system has monitored such systems for a time long enough to have observed a number of failure events, including all significant kinds of failure modes, that data can be used as a basis for predicting the remaining life of another system. Unfortunately, if the other system differs substantially from those used to generate the failure data, either in operating environment or in component details, the accuracy of failure prediction will suffer. Furthermore, such an approach cannot be applied to the first few units of a machine placed into service, as there is no prior failure data.

(2) Model-based. A "first principles" dynamic model of a mechanical system can, in principle, be developed and used to explore the response of the system to any contemplated operating environment. Challenges include anticipating significant system failure modes, including all the relevant physics, and running quickly enough to be used in near-real time. Clearly, the availability of such models would also be useful during the design of new machinery.

History-based prognostics is currently the dominant paradigm for condition-based maintenance approaches. Model-based prognostics is a growing area of contemporary research with clear, if longer-term, payoffs.

Development of efficient, low-order numerical models of material damage accumulation and component failure are key parts of a model-based prognostics capability. Because a machine is monitored essentially continuously, predictive models need not approach perfect fidelity, as model errors can be detected and corrected. On the other hand, such models must be accurate enough to be able to guarantee safe operation over some time window into the future.

HIERARCHICAL MODELING: A hierarchical modeling approach is being developed to address this need, as depicted in Figure 1. The modeling approach is motivated by recognizing that a mechanical system can be considered a hierarchical collection of interacting parts, and that system failure is typically the end result of a sequence of events stemming from an initial material failure. The modeling approach is complicated by the recognition that material failure involves complex, imperfectly understood processes.

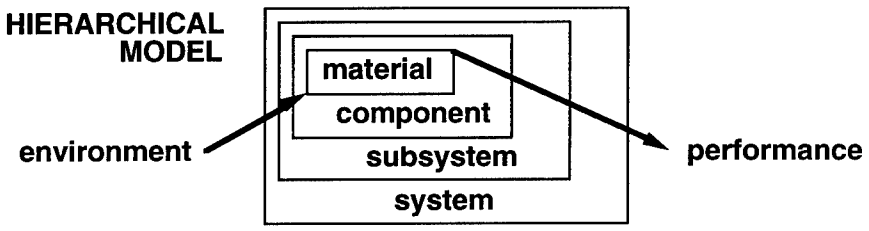


Figure 1: A Hierarchical Modeling Framework

The hierarchical modeling approach of present interest has the following features:

(1) Local material failure is modeled *approximately* using continuum damage mechanics;

(2) Component failure is modeled through the numerical solution of representative boundary value problems, permitting the identification of component "damage modes;" and

(3) System failure is modeled by representing individual components as discrete elements, with damage states associated with each component damage mode.

Additional details associated with each of these levels of failure are described in the following subsections.

Material Failure (Damage Mechanics Models): Failure by crack initiation and growth is one of the most important type of component failure modes anticipated. Although this might suggest a material failure approach based on detailed fracture mechanics modeling, a continuum damage mechanics approach is being pursued for two main reasons: First, the system-level models envisioned will employ specific component failure modes and associated damage states -- and continuum damage mechanics models are already in the desired coupled-field form. While fracture mechanics models could be used in principle to identify failure modes associated with cracks, parameterization of crack growth in a state variable model seems a formidable task. Second, fracture mechanics approaches appear to be inconsistent with the goal of efficient, low-order computational models. This said, research on fracture mechanics is certainly relevant, but not has been pursued as a key element of the present approach.

Researchers have been developing the largely empirical field of continuum damage mechanics for roughly 40 years, with the primary goal being the description of local material failure prior to the development of "macro-cracks." Researchers have considered brittle, ductile, creep, and fatigue damage (to name a few), and dynamic aspects including initiation, growth, and coalescence.

A continuum damage mechanics formulation involves the joint evolution in time of (at least) two fields: the mechanical displacement field and a damage field. Damage (D) is thought to degrade the elastic properties of the material (E); in a representative form, as follows:

$$\tilde{E} = (1 - D)\, E \tag{1}$$

And a typical damage evolution equation has the following form:

$$\dot{D} = H(\sigma) \frac{\sigma^p}{[A(1 - D)]^q} \tag{2}$$

The requirements for thermodynamic consistency in constitutive and evolution equations are well understood. Although researchers recognize the need to consider interactions among the various physical processes that may be operative, the presence of multiple degradation processes reduces the fidelity of model predictions. The availability of a machinery monitoring system for model error detection makes it possible to consider continuum damage mechanics for a model-based prognostics approach.

Component failure by crack growth may perhaps be captured to first order using such a damage mechanics scheme, by identifying a crack as a connected region of high localized damage. A companion research effort led to a new continuum damage mechanics model based on the notion of equivalent elliptical microcracks [5]. This model, illustrated in Figure 2, is distinguished by the following combination of features:

(1) A scalar damage field. This is defined as the local average volume fraction of damaged material. The orientation and aspect ratio of the equivalent elliptical microcracks are used as auxiliary fields. The orientation of the damage affects the

damage growth direction, while the aspect ratio affects the damage growth rate for a given stress level.

(2) Non-isotropic elastic behavior of damage material. Damage can fundamentally change the nature of material behavior from isotropic to orthotropic. Increasing damage tends to reduce the stiffness of the material, and the elliptical crack paradigm makes such reduction direction-dependent, perhaps capturing a key aspect of crack growth. The material fails locally when the matrix of elastic constants for the damaged material becomes singular, *i.e.*, when the material becomes incapable of carrying load in some direction.

(3) Empirical material-based evolution equation. The evolution equation for the damage field is motivated by empirical fracture mechanics relations. Damage growth is driven by a scalar damage equivalence stress exceeding a threshold value. The values of the auxiliary fields (orientation and aspect ratio of the equivalent elliptical microcracks) are determined from the current displacement and damage response.

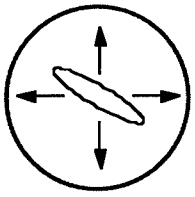


Figure 2: Equivalent Elliptical Microcrack as the basis for a CDM theory

Issues related to this specific damage mechanics approach to material failure include: evaluation of alternative methods for determining the current crack orientation; evaluation of alternative growth laws involving complex stress; and determination of the necessary material properties to be used in system simulations. These issues are being pursued in parallel with component and system modeling research.

Component Failure (Numerical Models and Damage Modes): Solutions of representative coupled-field boundary value problems for individual components continue to be pursued. The complexity of the coupled-field constitutive and evolution equations, not to mention the geometry of realistic components, necessitates a numerical approach. This is being addressed in the following ways:

(1) Demonstrating the proposed methodology through coupled-field finite element analysis initially, to be followed by the development and use of "meshless" numerical analysis to investigate spatial convergence of numerical solutions.

(2) Performing appropriate experiments to characterize material behavior and to determine model parameters. After these have been established, predictions based on the damage mechanics model will be compared to those based on fracture mechanics.

(3) Developing a methodology for identifying and describing component failure modes, based on a characteristic spatial distribution of damage.

Initial numerical analyses implementing our damage mechanics model have been accomplished using the commercial finite element code ABAQUS with specialized subroutines. Solutions obtained to date exhibit monotonically increasing damage accumulation to failure. Figure 3 shows damage growth in the vicinity of a hole in a bar subjected to uniform longitudinal tension.

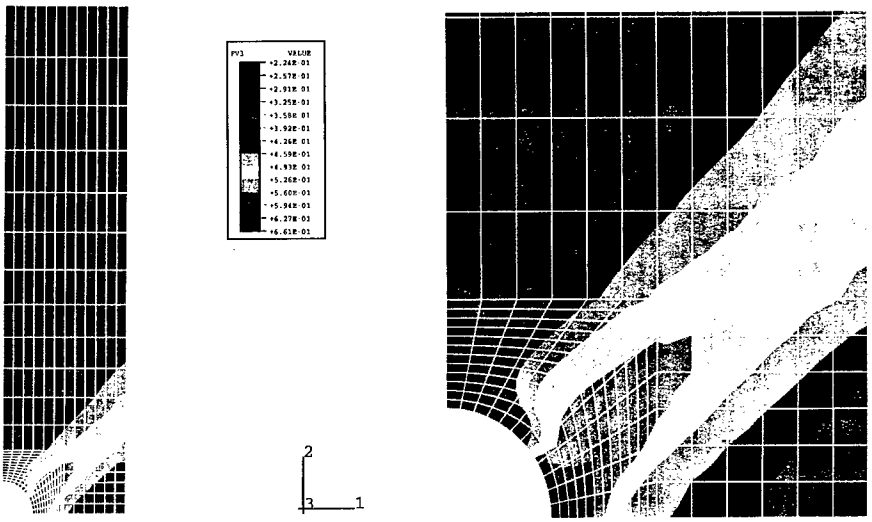


Figure 3: Damage Growth Near a Hole in a Bar Under Uniaxial Tension

The use of the Von Mises (deviatoric) stress and the "damage equivalence stress" (which includes a hydrostatic component) have been evaluated as alternatives for driving damage evolution. Damage evolution in a bar with an initial crack appears qualitatively similar to crack growth when the damage equivalence stress is used, and is more like a shear band when the Von Mises stress is used. Figure 4 shows representative stresses in two meshing planar gears. These stresses drive damage growth, especially in the vicinity of contact regions and near tooth roots.

Solutions obtained to date exhibit some mesh sensitivity. This is the subject of continuing investigation, with emphasis on developing meshless numerical methods. Such methods will allow additional degrees of freedom to be added to the model in regions of high response gradients without re-meshing. In addition, predictions made using the subject damage mechanics model are being compared to those obtained using conventional fracture mechanics methods. Obtaining a favorable comparison requires the use of appropriate material properties and parameters in the models.

Numerical results such as those shown in Figures 3 and 4 are being used to define *"component damage modes."* Such modes describe characteristic (or representative) spatial distributions of damage under expected operating conditions. Component health can

then be described in terms of the amount or degree of each damage mode present in the current response. This concept is fundamental to developing low-order system dynamic models.

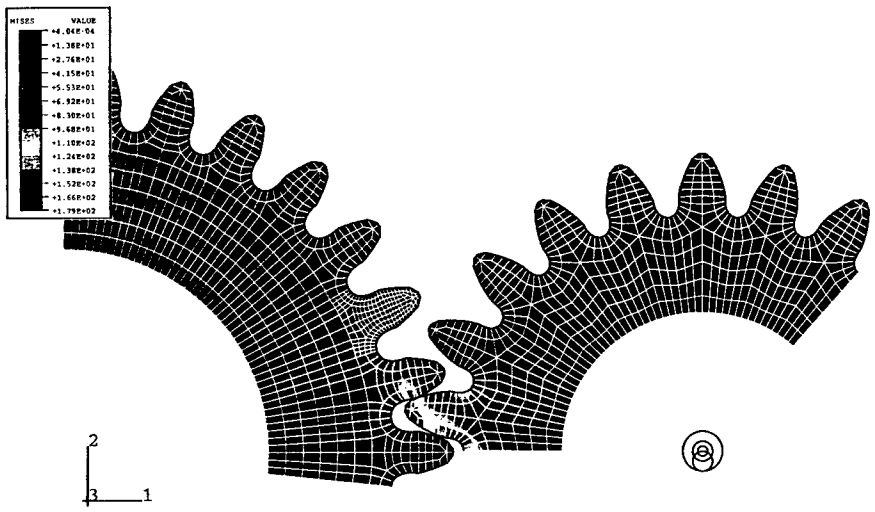


Figure 4: Stresses Drive Damage Growth in Meshing Gears

Figure 5 illustrates the description of gear tooth failure using an effective tooth stiffness over time. In this example, a damaged region growing from the root can be considered a significant damage mode, and the simplest discrete model of a tooth is just a scalar spring, the stiffness of which decreases with increasing damage.

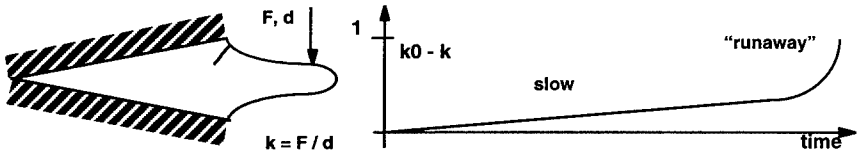


Figure 5: Failure of Single Gear Tooth Due to Root Crack

System Failure (Integrated Low-Order Component Models). Component damage modes will be used in the construction of low-order coupled-field component models for use in system simulation. These system simulations will be used to identify likely *"system failure modes,"* to anticipate accompanying observables, and to assess likely remaining life.

Component damage evolution will be described by a combination of scalar state variables and associated spatial damage distribution. The spatial distribution is likely to depend on the state variable in a non-linear way, unlike a conventional assumed-modes method. Figure 6 illustrates the concept of developing a discrete component model, including damage, from a continuous model.

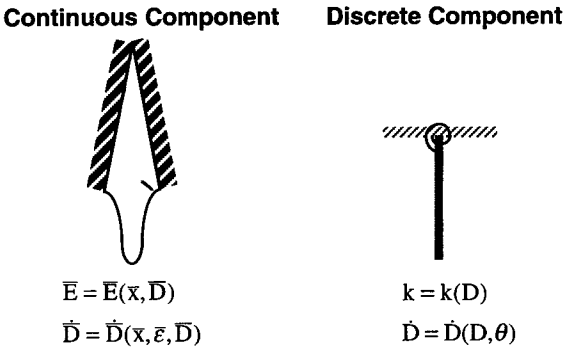


Figure 6: Discrete Component Damage Models are developed from Continuous Damage Models

Component models (discrete) will be assembled into system models, as illustrated in Figure 7. These system models will be used to identify likely "system failure modes," which involve a material failure that leads to a fundamental component failure followed by additional component failures and loss of system performance. The system models will also be used to identify accompanying observable signals, and to assess likely remaining life.

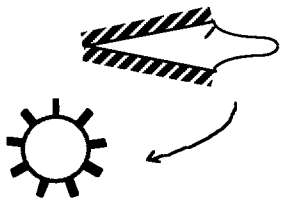


Figure 7: Discrete Component and System Models are developed from Component Models

SUMMARY. In the near term, it appears that history-based prognostics will dominate in practical applications. In the longer term, however, model-based approaches are likely to contribute significantly to prognostics capability, especially when little data exist for a particular machine or operating environment. Furthermore, a model-based capability is essential for the improved design of new machinery that is to be monitored and for training of maintenance engineers.

The hierarchical approach to model-based prognostics described in this paper is unified by the use of damage states at every level, states that are assumed to affect apparent stiffnesses and damping prior to failure. The approach has the following general features: (1) local material failure is modeled approximately using continuum damage mechanics; (2) component failure is modeled through numerical solution of representative boundary value problems; and (3) system failure is modeled dynamically by representing individual components as discrete elements having associated damage states. Damage mechanics solutions for individual components are used to define characteristic "component damage modes" that are used to construct low-order component models for use in system simulation. These simulations are used to identify system failure modes and associated observables, and to assess likely remaining life by predicting the consequences of continued operation.

Such an approach appears promising because the models used need not approach perfect fidelity (as model errors can be corrected by monitoring), but must only be accurate enough to guarantee safe operation over some limited time window into the future. Clearly, the successful deployment of this approach will not relieve engineers of the responsibility of judgment: anticipating elemental failure events and including the corresponding damage states in the system dynamic model. Anticipating such failure modes in advance, especially for new designs, will continue to be a challenge. Perhaps tools based on approaches like that described herein will enable that capability.

REFERENCES

1. Hansen, R.J., D. L. Hall, and S. K. Kurtz, "A New Approach to the Challenge of Machinery Prognostics," *Journal of Engineering for Gas Turbines and Power*, Vol. 117, No. 2, April 1995, pp. 320-325.

2. Rosenberg, Bruce J., "The Navy IDSS Program: Adaptive Diagnostics and Feedback Analysis - Precursors to a Fault Prognostics Capability," *IEEE Proceedings of the National Aerospace and Electronics Conference*, Vol. 3, 1989, pp. 1334-1338.

3. Parker, B.E. Jr., "Helicopter Gearbox Diagnostics and Prognostics Using Vibration Signature Analysis," *SPIE Proceedings*, Vol. 1965, 1993, pp. 531-542.

4. Ray, Asok, "Stochastic Modeling of Fatigue Crack Dynamics For On-line Failure Prognostics," *IEEE Transactions on Control Systems Technology*, Vol. 4, No. 4, July 1996, pp. 443-451.

5. Lee, Usik, G.A. Lesieutre, and L. Fang, "Anisotropic Damage Mechanics Based On Strain Energy Equivalence and Effective Elliptical Microcracks," *International Journal of Solids and Structures*, to appear, 1997.

Developing Prognostics For Diesel Engines

E J Lewandowski
Cummins Engine Co.
P.O. Box 3005
Columbus, Indiana 47202
e.j.lewandowski@cummins.com

Abstract: Developing prognostics for diesel engines offers significant benefits to the end user. This technology can increase equipment utilization, reduce operating costs, and increase system reliability (measured in uptime) without having to significantly increase component durability or reliability. A management system which incorporates prognostics is described. Research needs related to developing prognostics are identified. Trending based on data gathered over time is discussed, along with different methods for gathering trending data on diesel engines.

Key words: Blowby, data gathering methods, diesel engines, prognostics, trending

INTRODUCTION: Much research and development has been done in the areas of engine diagnostics and maintenance. Relatively little progress, though, has been made in developing prognostic technology, which predicts failures, optimizes maintenance, and quantifies remaining life [1]. Prognostic technology offers significant benefits to the engine user. This technology can increase equipment utilization, reduce downtime, reduce operating costs, and increase system reliability without having to greatly increase component durability or reliability. The possibilities in theory are prodigious; the fact is that the industry is years away from commercializing these methods.

This paper focuses on prognostics for the commercial diesel engine, although the concepts could apply to internal combustion engines in general. The framework for diesel engine prognostics is similar to that for prognostics or condition-based maintenance of other equipment. This paper highlights some of the special aspects of diesel engine prognostics.

Role of diagnostics and prognostics in improving engine reliability: Customer demand for ever-improving product reliability does not mean that engines will become reliable to the point that diagnostics and prognostics will be unnecessary. Rather, there is a cost-driven balance between improved engine reliability and diagnostic and prognostic capability. A component-by-component beef-up approach to improving reliability will probably yield an engine which is too costly and heavy. The appropriate approach may be to continue beef up certain components and subsystems, combined with adding diagnostics and prognostics. A less reliable engine can become reliable from the end user point of view with the addition of effective diagnostics and prognostics.

EQUIPMENT MANAGEMENT SYSTEM INCORPORATING PROGNOSTICS

Prognostics must be understood within the context of a complete system which helps the end user better utilize his equipment. The term 'management' is used to point out that prognostics is not an end to itself. Rather, prognostics serves to support equipment management. While a complete equipment management system includes diagnostics and many other components, the system described here focuses primarily on the prognostics portion.

Figure 1 shows what the prognostic portion of an equipment management system might look like. The first major part of the system is to identify and characterize the health of the engine. The next step is to prognosticate about engine and component performance. In conjunction, one must optimize maintenance and repair based on the prognostic information. Finally, this information is converted into a concrete form which the end user or control system can act on. The various parts of this system are discussed in detail next.

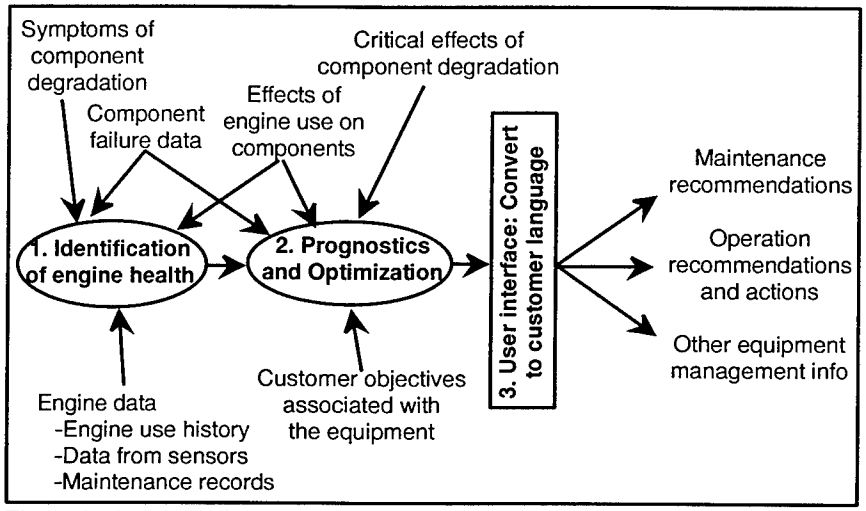


Figure 1: A prognostic management system.

Understanding Component Failure: Developing prognostic capability requires improving our understanding of component failure. Today failure mode and effect analyses (FMEAs) are done on most engine components and subsystems. These FMEAs identify potential failure modes and severity, and are used to design out higher risk failures. A prognostic system focuses on those failure modes which are statistically significant and actually occur in the field. The FMEA methodology can be adapted to identify where prognostics would be most beneficial.

The nature of the transition from a functional component to a failed component is important. Many mechanical components degrade over time through mechanisms of wear,

creep, corrosion, and other microstructural changes [2]. The rate of component degradation versus time needs to be understood. For example, does a component initially wear slowly, followed by a period of rapid wear to complete failure? Or does wear occur gradually, almost linearly, over time, until complete failure?

Historical failure data can provide a way to predict remaining life for come components. This data needs to be complete enough to accurately characterize the variability in component life - i.e., the type of distribution, range, etc.

Component failure information such as this will be used to predict when to service the engine and will determine what thresholds to use for various decision points [3].

Predicting Impending Failure Based on Use: Much maintenance done today is based the number of hours or miles of use. Filter changes, oil and coolant changes, even engine overhauls are based on use. This method needs to become more sophisticated, since some maintenance is done more often than needed while other maintenance may not be done often enough. Components which already are maintained based on use need improved methods for calculating remaining life. This method must be expanded to cover other components as well. Better understanding the effect of use history or duty cycle (stress, temperature of operation, etc.) on component life can noticeably reduce the life variability seen in specimens of a particular component. This component level use information can then be related to engine duty cycle (torque vs. speed vs. time), number of engine starts, number of cold starts, time spent operating very hot or very cold, the condition of the lubrication used, and so on to estimate component life.

Identifying Component Degradation and Predicting Impending Failure Based on Symptoms: As a component degrades, various symptoms will appear. These symptoms need to be identified, and they need to be correlated with degrees of degradation. Practical considerations with diesel engines suggest the focus should be on identifying macroscopic symptoms detected by speed, torque, pressure, temperature and other sensors, and by oil analysis and periodic engine health checks. Identifying microscopic symptoms requiring partial engine disassembly and metallurgical analysis are generally not appropriate in this field.

Critical Effects of Component Degradation and Failure at the Overall System Level: Certain effects or symptoms of component degradation are especially critical in the commercial engine market, and merit special attention. These symptoms are listed below.

Loss in fuel economy: The degradation of certain components affects fuel economy. The relationship between component degradation and fuel economy is approached from two perspectives. First, fuel economy can be a symptom or measure of component degradation, useful in quantifying engine health and predicting remaining life. Second, the end user has certain fuel economy targets based on operating costs. The potential loss in fuel economy can trigger maintenance recommendations and corrective actions.

Loss in performance: Certain components impact important end user performance requirements of response, power output, and stability. As with fuel economy, performance is a symptom of engine health and an end user objective. Maintenance action may be triggered based a user's allowable degradation in a certain critical performance parameter.

Increase in chemical emissions: The effect of component degradation and failure on emissions is taking on a greater and greater role in the overall exhaust emissions control strategy. Research has shown that engines and vehicles with significantly degraded components can contribute to exhaust emissions increases. To this end, the California Air Resources Board (CARB), which sets emission-related standards in California, and the federal Environmental Protection Agency (EPA) have set standards relating to diagnostics and emissions levels. Today manufacturers of certain vehicles must provide on-board diagnostic systems which are able to detect the degradation of certain systems which cause an increase in emissions in the applicable Federal Test Procedure (FTP) or in any reasonable in-use driving condition [4]. As a result, the relationships between component degradation and emissions increase are already being studied. This is an important part of the engine maintenance equation.

Degradation causing collateral damage: If the degradation or failure of a component causes serious damage to the rest of the engine, clearly that part must be repaired sooner. Less risk is allowed in delaying the replacement of that component.

Optimizing Utilization Based on End User Objectives: With information on component life and effects of component degradation and failure, the next step is to optimize utilization of the engine system. From the end user perspective, cost needs to be considered. Extending the life of the engine at any cost does not optimize utilization from the end user perspective. All costs directly and indirectly related to the engine as a component of the end user's business system must be factored in. This means that costs related to downtime should influence the recommendations of a prognostic management system. The trade-off between delaying a repair to reduce cost versus increasing the risk of downtime is factored in. Optimizing maintenance scheduling is another part of the equation. For example, an engine which is down for major repair which is expected to require a minor repair soon maybe should have the two repairs done at the same time.

Quantities like the cost of downtime, the value of reliability, the value of fuel economy, the value of emissions can vary from one engine user to the next. The prognostic management system must be able to take into account each user's priorities in making maintenance recommendations. For example, users operating standby generators have much less concern for fuel economy compared to those running prime power generators. Reliability, on the other hand, is extremely important. Repairs which enhance fuel economy with little or no effect on reliability would have a lower weighting. Users operating generator sets as main sources of power will want the prognosticator to tell them when fuel economy has degraded even slightly and what to do about it. It is important that the system be flexible and adaptable.

Clearly, factoring in all of these considerations is not a trivial problem. The fact that these constraints and data exist in several different domains (OEM data, engine data, customer business objectives, maintenance schedules, etc) further complicates the problem.

The Interface With the User: The prognostication and optimization process may generate information related to the likelihood of various failures at some future time frame. As such, this information is of minimal value to the end user. It must be translated into a form which can help decide things such as "when should the engine be overhauled?" and "what repairs need to be made at what time?" Maintenance recommendations should blend in with the user's complete equipment management system. Information should be provided at the equipment manager's request. The equipment manager should be able to query the system about equipment condition, needed repairs, downtime, etc., and get a clear picture he can understand.

Feeding Back Equipment Operation Recommendations and Actions: The prognostic system not only helps maintain and repair an engine, it can also adapt how the engine is operated in order to better utilize it. One could modify engine operation so as to limit stress to a component with minimal life left which the user wants to last a certain length of time. The system could compensate for performance changes at the request of the user. It could prevent the operation of the engine in certain ways which are detrimental to long engine life when the user values engine life as a high priority and the engine is operated in a detrimental way frequently [5, 6].

TRENDING

Trending is a somewhat heuristic prognostic method which is beginning to be used with diesel engines. One looks at trends in data from individual sensors over time, then makes decisions based on how things are changing relative to certain limits[7, 8]. This section discusses some of the issues around trending with engines.

Blowby and Trending: Blowby is the combustion gases which leak by the piston rings, exhaust valve stems, and turbocharger bearings. As certain engine components wear (esp. rings and liners), blowby increases. While blowby is a good indicator of engine health, blowby in and of itself is not sufficient to recommend a specific course of action or estimate remaining life. High blowby can relate to the need to overhaul an engine; it can also indicate the need to address a turbocharger issue. Additional sensor or diagnostic information is necessary to assess engine and component health. Being able to make this assessment with the current suite of sensors and without extensive diagnostic analysis is important to making the system work cost effectively.

Even having identified where the blowby is coming from, there is little quantitative information available relating a particular blowby level to the need to perform a repair. Many blowby guidelines published by manufacturers say when blowby levels will potentially cause damage to the engine, not when levels indicate a problem needing repair.

From a predictive point of view, little work has been done to relate blowby over time to component deterioration.

Blowby is a good example of the potential of trending data for diesel engines. Many other pieces of data offer similar potential in helping the operator manage his equipment. The same sort of research needs to go on with this data as well, making connections between component degradation and failure and changes in parameters.

Data gathering strategies for trending: Today, data can be gathered faster and in much greater amounts than it can be stored, managed, and processed. If we had unlimited memory and data storage, transmission, and handling capability, we could store all data continuously. This would give us a complete picture of the engine history and operation. That is not realistic today, nor will it be realistic any time in the near future. Therefore, one must be selective in the initial gathering of the data. Data may have to be sorted and compressed after being gathered. In this section we will discuss six possible ways to gather data for trending. They are:

1. Continuous datalogging - very fast sampling
2. Datalogging at rated power
3. Datalogging with variable load/speed thresholds
4. Datalogging over loadblocks
5. Continuous datalogging - periodic sampling
6. Other datalogging methods

1. Continuous datalogging - very fast sampling: This is the ideal datalogging method mentioned above, where data is logged continuously regardless of operating condition. This means that during engine operation, data will be taken on a periodic basis at perhaps once per second. Data would be logged no matter how long or short the engine operated at a particular condition.

The advantages of this method are that no matter where the engine operates, one will get data. One can do some things which one cannot do with some of the other methods. When looking at the data, one can determine whether the engine is at a steady-state or transient operating condition. One can also see exactly how the engine is being used. (Actually, to really see events like throttle snaps, one should sample at around 20 Hz to get an accurate picture of the event.) One can accurately make duty cycle calculations - (what percent of the time the engine spends in various operating areas on the torque-speed map).

The disadvantage of this method, as mentioned before, is that it simply is not feasible. The rest of the data gathering methods are various ways to overcome this issue.

2. Datalogging at rated power: Taking data just at rated power is appealing for a number of reasons. Much engine test data is generated at rated power. This makes it easy to compare actual engine performance with expected performance. The engine speed and torque are fairly high, resulting in generally high temperatures and pressures throughout

the engine. This improves the signal-to-noise ratio, making changes in data easier to detect. Variability is low because rated power is only one operating point, not a region.

There are some drawbacks with datalogging at rated power. Not all applications always operate at rated power on a regular basis. This can be because the engine's load—a hydrostatic pump, a torque converter, a transmission—is set up in such a way that the engine can never get to rated power. Or the operator may find that he never needs full power for the particular application, so he seldom runs up to rated.

Another question is what constitutes rated power? Does the engine need to be within 1% of rated power? Is 5% or 10% sufficiently close? Broadening the band around rated power increases the likelihood that data will be gathered and that there will be more of it. However, as the band broadens, the variability in sensor data increases. Trends are more difficult to pick out.

There are some trends in engine parameters which may show up better at different operating conditions - perhaps low idle, high idle, or torque peak. Taking data at rated power may not capture information which would be useful for later analysis.

Obtaining steady-state data - There is a trade-off between obtaining stable steady-state data and obtaining sufficient data. When taking data for trending purposes, one would generally want to be operating the engine at a steady-state condition. Some parameters like fuel pressure stabilize quickly. Temperatures often take longer because of lags both in the engine and in the sensors themselves. For example, exhaust temperature sensors can take 2 minutes to stabilize, and they still may continue to drift a few degrees for several minutes. The longer one waits, the more consistent the data will be. But if one waits too long, the engine may leave that particular operating point. Some applications, such as front end loaders, may seldom operate the engine at one torque-speed condition for even two minutes.

3. Datalogging with variable load/speed thresholds: To get around the problem where not all applications operate regularly at rated, one could make the load and speed threshold above which data are logged calibratible. When the engine is installed or set up, the power and speed thresholds can be set to appropriate levels for the application. If a particular application is limited to a speed below rated, then the speed threshold would then be set accordingly.

This method suffers from the inconvenience added by having to uniquely calibrate load and speed thresholds for different applications. One could make the software find these thresholds adaptively. The software could look to see what is the highest speed and load at which the engine operates, and then adjust the thresholds accordingly.

This method introduces much variability into the data. From application to application, the pressures, temperatures, and speeds will be different in both their mean and their deviations. It will be difficult to set firm guidelines for expected values for all engine applications, because of variation in the engine operating point.

4. Datalogging over loadblocks: Another datalogging method is based on putting data in loadblocks. This means that the torque-speed map is broken up into many blocks. If an engine never operates at rated, the neighboring blocks will likely have data which can be used for analysis. If a trending algorithm requires data at a certain load or speed, the algorithm can go to that block and use the data there.

This algorithm can be implemented whereby the entire torque-speed map is used, or just portions of it. The difference between this method and continuous data logging is that the data would be taken only when the engine is at quasi-steady-state conditions.

5. Continuous datalogging - periodic sampling: Rather than logging data every second or so, one could log data every minute, every 5 minutes, or every hour. The amount of data would go down by a couple orders of magnitude, simplifying data handling.

Some information is lost in the process of reducing the sampling rate. One no longer has any idea if the engine is in a steady-state mode when a data point is taken. This will increase variability.

6. Other datalogging methods: There are other combinations and permutations of the above methods which could be used. In addition, one could employ some other schemes:

- Adaptively modify the datalogging scheme depending on how the engine is operated.
- Attach information to the data taken at a particular instant noting how long it has been operating near a particular condition before the data was taken.
- Record lots of data continuously, then search for certain predetermined operating conditions. For example, the datalogger may look for a throttle snap, steady-state operation at rated, steady-state operation at idle, high idle operation, engine start-up, engine shut-down and so on. The datalogger will then save just these pieces of data, acting as a prefilter. The approach here is to save good data from just a few conditions.

Engine modeling and trending: Engine models can be used to get around some of the problems in data gathering. Engine models can be used to 'normalize' data. Much of the manufacturer's engine data is taken at very controlled conditions. The data from engines in the field, for the most part, will almost never be taken at those same conditions. An engine model can be used to take the data from the field and extrapolate/interpolate/convert the data to what they would be at some standard engine operating condition.

An engine model also could be used to 'normalize' the trending data taken at a given speed and load for environmental changes. The engine operating conditions will vary from hour to hour, day to day, season to season, and so on. A model could be used to correct for changes in altitude, intake air temperature, and other critical environmental parameters. This should reduce variation in the data, making it easier to pick out the real trends in the data.

SUMMARY: This paper has identified many of the areas of research and development to bring prognostics to diesel engines. It has detailed the larger management system which includes prognostics. It is important to consider the end user's needs and requirements as part of this system in order for prognostic technology to succeed. Trending, a first step towards developing prognostics, is starting to be used. Determining what data to gather and how to gather it becomes very important to making trending work. The next several years should see many of these ideas taking form as research continues.

---

[1] G.W. Nickerson and D.L. Hall, "Research Imperatives for Condition-Based Maintenance," COMADEM '95.

[2]Robert J. Klova and Hermes Galatis, "Power Turbine Remaining Life Analysis: Experience and Practical Considerations," Presented at the International Gas Turbine and Aeroengine Congress & Exhibition, Birmingham, UK, June 10-13, 1996, ASME paper 96-GT-211.

[3] J.H.J. Geurts, "Optimal Age Replacement Versus Condition Based Replacement: Some Theoretical and Practical Considerations," *Journal of Quality Technology*, Vol. 15 No. 4, Oct. 1983, pp. 171-179.

[4] California Code of Regulations, Section 1968.1 of Title 13.

[5] A. Ray, M-K. Wu, M. Carpino, and C.F. Lorenzo, "Damage-Mitigating Control of Mechanical Systems: Part II - Formulation of an Optimal Policy and Simulation," *Proceedings of the 1993 American Control Conference*, pp. 3146-3148.

[6] W.C. Merrill, J.L Musgrave, and T.H. Guo, "Integrated Health Monitoring and Controls for Rocket Engines," SAE Paper No. 921031, 1992.

[7] L.B. Peterson, "Trending and Data Acquisition System Monitors Single or Multi-Engine Plants," *Diesel & Gas Turbine Worldwide*, Vol. 22, No. 6, July-Aug. 1990, pp. 10,12.

[8] S.C. Hadden, and J.A. Maurer, "Predicting Failure Using an On-Board Monitoring System," *Proceedings of the 35th Meeting of the Mechanical Failures Prevention Group*, Gaithersburg, MD, Apr. 20-22, 1982, pp. 229-236.

INTELLIGENT ON-LINE DETECTION OF CAVITATION IN HYDRAULIC PUMPS

John P. H. Steele
Laboratory for Intelligent Automated Systems (LIAS)
Center for Robotics and Intelligent Systems
Colorado School of Mines
Golden, CO 80401

Abstract: Pump cavitation can have devastating effects on the pump in which it occurs. High pressure hydraulic pumps, the type used to provide the power for heavy equipment and robots, can be completely destroyed if they are allowed to cavitate for even a few hours. Similar results are true for feedwater pumps in steam power applications and hydraulic power units for aircraft controls. While cavitation is obvious to any mechanic standing near a pump which is cavitating, it is difficult to diagnosis if the pump is being operated remotely or is just one piece of equipment among many. The goal of this work is to develop an on-line condition monitoring tool to detect the occurrence of cavitation in hydraulic pumps using expert system technology, and to communicate that condition to operators and management using an intelligent user interface. The method of cavitation detect presented here involves developing a linear discriminant function of five variables. These five variables are state variables which would normally be available in many applications, i.e., pressures, temperatures, and flow rate. This approach is shown to be successful from tests run on a hydraulic system. In addition, the use of on-line cavitation detection is presented in the context of it being only one of several system attributes monitored as part of an integrated real-time machine health assessment system (condition monitoring). The goal of the system is to provide on-line intelligent predictive maintenance capability thus saving time and money on mission critical systems.

Keywords: Condition Monitoring; Pump Cavitation; Online Diagnosis; Expert Systems

Background: Pump Cavitation is a significant problem. In hydraulic powered equipment, if the HPU (Hydraulic Power Unit) pump cavitates, for even a few hours, the pump will be destroyed and the system will have to be removed from service for repair. For mission critical equipment this can have important financial implications.
While the effects of cavitation are well know, the actual behavior and its relationship to different pump types is less well understood. For example, Arnold (93) has characterized cavitation as being either suction cavitation or discharge cavitation. Suction cavitation is caused by high suction pressure (vacuum). Discharge cavitation occurs when the discharge head is too high. The cavitation takes place in the discharge area of the pump. Mackay (93) looked at vane pumps and categorized the failure modes into four conditions: cavitation, air entrainment, suction recirculation and discharge recirculation.

These differ in the location within the pump where they occur. In addition, the vibration induced by cavitation can cause secondary failures in seals and bearings.

While these causes are still being investigated, the ability to detect cavitation is in little dispute. In the case of high pressure hydraulic pumps, the occurrence of cavitation is signaled by accompanying audible emissions which would be noticeable to even an untrained observer nearby, if the background noise level is not too high.

However, for many systems, the goal is to have the equipment working remotely from human operators and thus there is a need for sensor based cavitation detection, and to have this capability on-line so that an operator can be informed of the occurrence via an alarm when it happens so that appropriate response measures can be taken. For example, cavitation is a major concern for those who maintain high performance aircraft (Wolf and Ferrara 94).

Goals and Approach:

The goals of this project are:

1. To demonstrate the ability to monitor the activities of the pump of a Hydraulic Power Unit (HPU) *on-line*, using generally available sensor based data.
2. To gain greater understanding of the relationships among pressure, temperature, and flow rate in the occurrence of cavitation in the HPU pump.
3. To demonstrate the ability to use Artificial Intelligence (AI) technologies to *infer* system health state from this monitoring in *real-time*.

The approach taken to providing on-line machine health condition monitoring here is

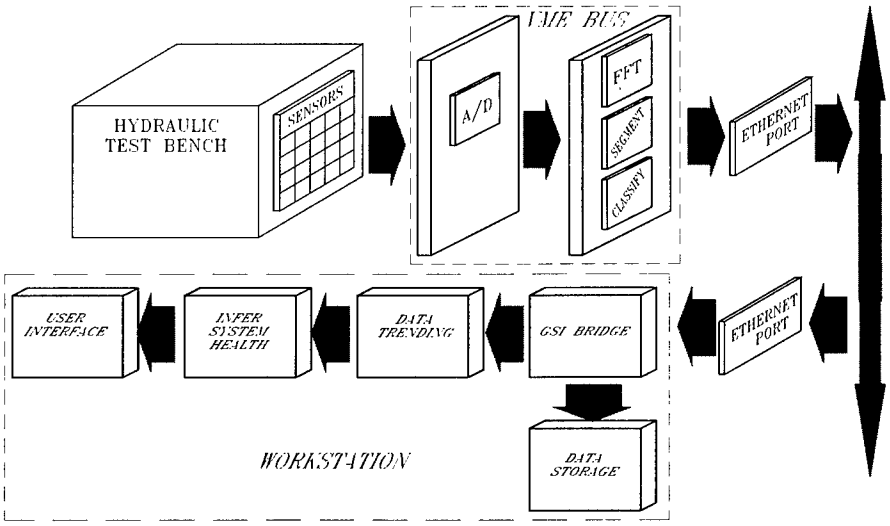


Figure 1: SHARP System Architecture

called **SHARP**, System Health Assessment and Real-time Prediction. This technology is intended to be a valuable tool for those tasked with maintaining high cost equipment and machinery in a variety of industries. The goal of this technology is to provide on-line monitoring of system behavior (machine condition) and provide assessment of the operational condition of the equipment to the operator and management in real-time, i.e., assessment based upon current measurements as well as historic trending information. In addition, the long term goal is to provide management with predictions of the remaining useful life of the machine before maintenance, repair, and/or replacement activities are required (predictive maintenance). This will allow management to plan and prepare for the removal of expensive and critical equipment from service rather than being forced to react to an unscheduled outage, or to pull fully functional equipment from service for scheduled maintenance when it is not really necessary.

System Architecture

The general architecture of SHARP is shown in Figure 1. The key features include the fact that the monitoring system is seen as a distributed system, i.e., sensors are located on the HPU, data acquisition and preliminary data reduction / abstraction are executed by a local processor associated with the monitored system. The reduced data/information is communicated via Ethernet to a client computer which contains the information interpretation software and information storage as well as the user interface (MMI, man machine interface in some industries).

In this work, pressure, temperature, and flow rate have been used as the sensor suite. While not used in this system as presented, accelerometers are also valuable sources of cavitation detection.

Data Acquisition is conducted using 12 bit single ended A/D conversion using an Acromag 9320 board. In SHARP, this data can be reduced and abstracted through a variety of techniques including Line Segment representation (King 95), Fast Fourier Transform, and simple classification. The resulting information is then shipped via Ethernet to the client computer. G2, a real-time expert system, obtains this information by using bridge processes written in C. The information is first archived for future reference before it is sent to routines which generate overall trending representations. Then either in the form of current

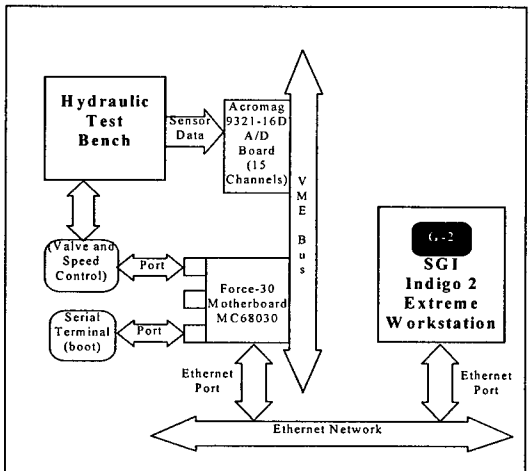


Figure 2: Physical Layout of SHARP Monitoring System

values, or historic trends, the information is provided to an inferencing engine which establishes system condition . These results are presented to the user via an object oriented schematic of the system . This interface can provide alarms when cavitation conditions are detected.

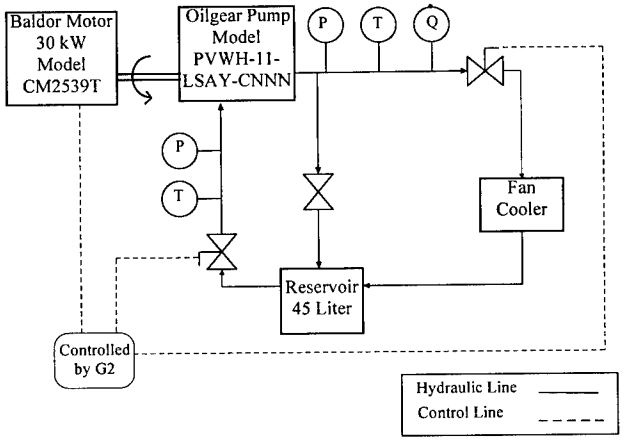


Figure 3: Hardware Schematic of Hydraulic Test Bench

A hardware representation of the system arrangement is shown in Figure 2. While only one server (the VME based 68030) is shown, the design is intended to be general enough to allow for several local servers for different equipment involved in a system to be interconnected.

To develop a detection system for cavitation in hydraulic pumps a hydraulic test bench has been constructed. The hardware schematic of this system is shown in Figure 3. Rather than have the load be implemented by an actuator, it is simulated by dropping the pressure across a load valve. The valve opening is adjusted to provide different load conditions. The inlet fluid flow to the pump is altered by means of an additional valve.

Table 1: Operational Parameter Matrix

Motor Speed [rpm]	Load Pressure (Output Impedance) [kPa (gage)]	Supply Pressure (Input Impedance) [kPa (atmosphere)]
750	0	101
900	2000	85
1050	4000	70
1200	8000	55
1350	12000	40
1500	16000	30 (or when cavitation first noted)
1800	20000	

Table 2: Test Parameter Matrix

Control Parameters	Measured Parameters
*COMMAND FREQUENCY, F	PUMP INPUT PRESSURE, Pi
*PUMP INPUT VALVE POSITION, Ri	PUMP OUTPUT PRESSURE, Po
*PUMP OUTPUT VALVE POSITION, Ro	PUMP INPUT TEMPERATURE, Ti
	PUMP OUTPUT TEMPERATURE, To
	PUMP OUTPUT VOLUME FLOW RATE, Q

By restricting the flow from the reservoir to the pump inlet, cavitation can be induced in the pump.

Experimental Setup: Inlet pressure, outlet pressure, inlet temperature, outlet temperature, and flow rate are all sensed. The temperature sensors are thermistors, the pressure sensors are diaphragm type with surface mounted strain gages, and the flow sensor is a turbine wheel with pick up. The signals were configured to be current based 4-20 milli-amp output to aid in noise suppression. Filtering was required of all signals because of PWM generated noise from the motor controller. Sampling rates were set at one kilohertz, but were averaged (see below). Data are stored on the Force 30 CPU until requested by the bridge process running on the SGI.

Test Procedure: The testing procedure involved the manipulation of three operator controlled variables: 1) command frequency (F) i.e., motor speed, 2) pump input hydraulic impedance (Ri) i.e., inlet valve setting, and 3) pump output hydraulic impedance (Ro), i.e., pump outlet valve setting. The operator-controlled variables were allowed to vary from minimum motor rotation (750 rpm), valves full open, to 100% (maximum motor rotation, 1800 rpm, valves closed), each in increments as shown in Table 1. For each of the resulting combinations of F, Ri, and Ro, a full set of SHARP attributes (Pi, Po, Ti, To, Q) were obtained in realtime and recorded. Observations by the operator were also recorded. The sequence was such that the supply pressure (vacuum) was the inner most loop and was varied with each data set taken. This was done to provide the best control over the cavitation state. The sequence was terminated whenever it was deemed that the cavitation state had been reached so as to minimize the actual number and duration of cavitation occurrences. "Hard" pump cavitation is easily identified acoustically. The matrix of operating points for the system are listed in Table 2. Note that data were taken at actual system settings which were representative of all combinations of variables, i.e., motor speed was the only setting which was matched exactly with the proposed settings. In some cases, cavitation was reached prior to obtaining the minimum supply pressure setting, Ri, and as a result, less than the 294 possible combinations were actually used.

Data Gathering: The data gathering process can be described as "For each speed, for each load pressure setting, for each inlet pressure setting, read all five sensors, two hundred fifty times." All measurements are assumed to be quasistatic, i.e., at each setting the system was allowed to stabilize prior to data gathering. At each setting, the data collection routine gathered 250 samples of each variable. This was done to provide statistical analysis of the data. These data were then processed to provide mean and standard deviation measures for each system setting.

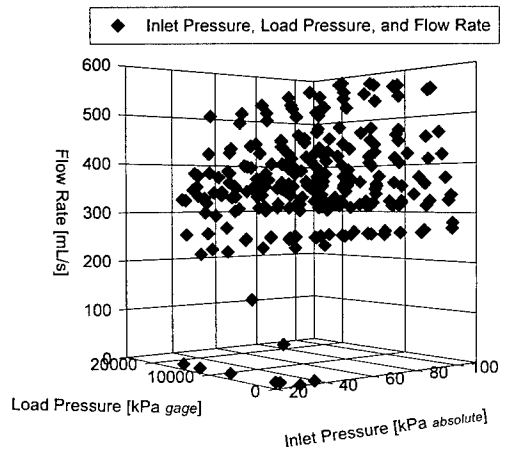


Figure 4: Test Parameter Results Matrix

Averaging techniques: Analysis was done after the testing to see what the minimum number of data samples necessary to obtain reasonable results for the mean and standard deviation would be. The results of this investigation indicated that while the mean value could be safely estimated from a hundred data samples, the standard deviation would only come into agreement after two hundred samples (Smiley 95). We have therefore set our sampling limit at two hundred samples per data point when gathering system state data. While not discussed here, noise was a significant challenge to this project because the motor driver is a high amperage PWM controller. Significant work was required to shield, ground, and filter the input data signals so that quality data could be obtained (Muzal 96).

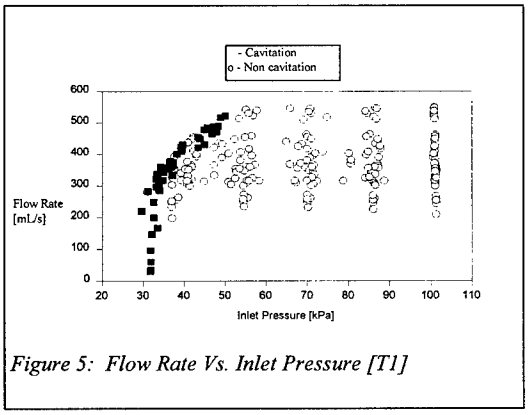


Figure 5: Flow Rate Vs. Inlet Pressure [T1]

RESULTS:

To display the space of parameters investigated, a plot of

the resulting mean values for each of the three independently varied parameters is shown in Figure 4. The plots of averaged data shown in Figures 5 & 6, represent different views of the data. Figure 5 is the plot of flow rate as a function of inlet pressure. The shaded square data points represent the system settings which resulted in cavitation. This graph clearly shows the relationship between inlet pressure and cavitation. However, one should note that in fact, cavitation occurs at different pressure settings depending on the other system setting, e.g., load pressure, temperature, etc. Thus it is not, in general, possible to state that at a given inlet pressure, cavitation will commence. It is also obvious that cavitation will occur at different flow rates, even at relative high values. Figure 6 shows flow rate as a function of load pressure (output pressure). This plot shows that cavitation can occur for any load and flow rate combination. Thus no configuration or operation of the robot would be exempt from possible cavitation.

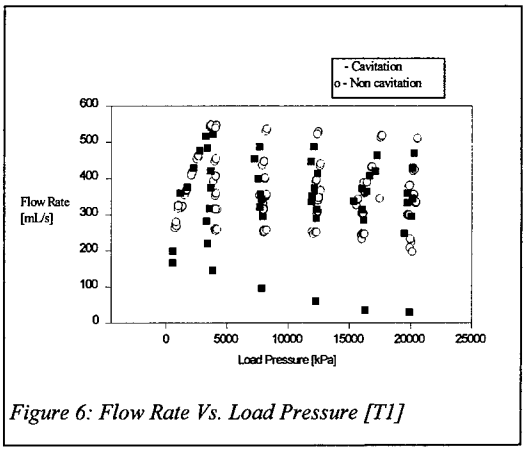


Figure 6: Flow Rate Vs. Load Pressure [T1]

Linear Discriminant Analysis

Relationship of Supply Pressure and Pump Motor Speed at Cavitation

As can be seen in Figure 7, as the rpm of the motor increases, the supply pressure at which cavitation occurs also goes up. Thus dynamic alarming is required for cavitation if a variable speed pump is used. Typically alarm limits are set statistically, and as a result they would not meet the need for condition monitoring at different speeds. For example, if a cavitation alarm were set to deal with motor speeds of 750 rpm, then the threshold might be a value of 40 kPa absolute. This would be fine until the motor speed was changed to 1800 rpm, since at this new speed cavitation could occur at much higher pressures.

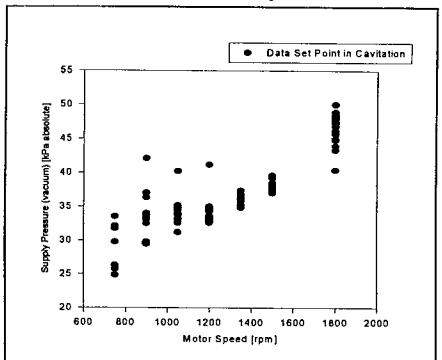


Figure 7: Low Pressure Vs. Motor Speed at Cavitation

Discrimination of Cavitation State: In previous work, Hooley (95) had shown that the cavitation state could be discerned from the data of the five principle variables (Pi, Po, Ti, To, Q). While this was first achieved by training a neural network to assess cavitation state from data sets, the analysis demonstrated that in fact the states of cavitation and non-cavitation were linear separable. Since the states were linearly separable, an approach call the Fisher Linear Discriminate, FLD, can be applied to the hydraulic pump data to develop a criterion by which the state of the pump can be determined. This process was performed on the two data sets which were developed from our testing. The results of that process are plotted in Figure 8. One can explain the process of developing the FLD as the process of seeking to find a plane in five dimensional hyperspace (note the importance of each variable is individually weighed) which allows the "best" clustering of the cavitating and non-cavitating data examples. Once this plane has been found, all of the data is then collapsed onto this plane, and the resulting clusters are displayed. A line drawn on the plane represents the best line of separation between the cavitating and non-cavitating clusters, see Figure 8. The correlation for this discrimination was 0.71. Coefficients for the first discriminant are presented in Table 3.

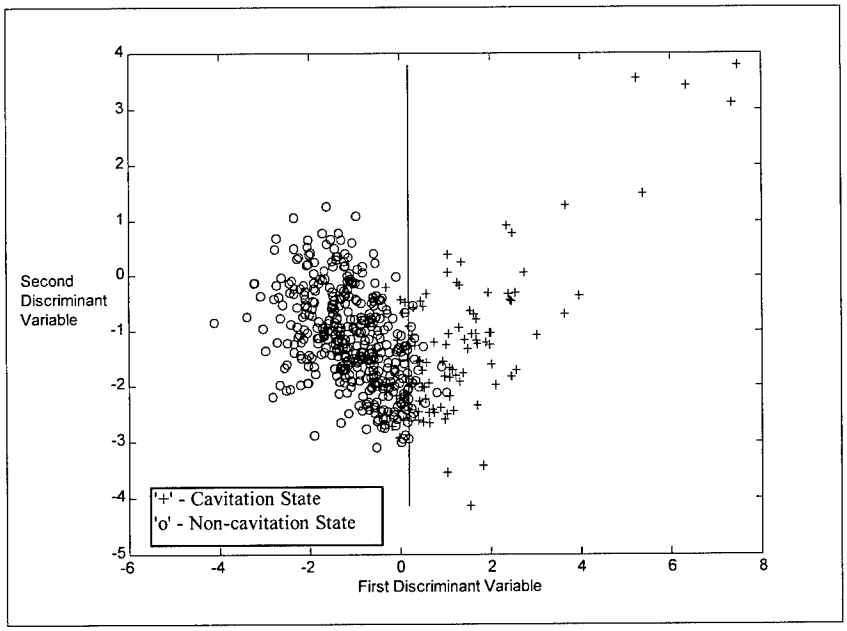


Figure 8: Linear Discriminate Results, T1 & T2 Combined Data

Using the "mean" criterion, i.e., finding the half way point between the two centroids of the two groups (non cavitation [o], and cavitation [+], the probability of committing a Type I error is 7% (predicting no cavitation, when in fact there is cavitation). The probability of a Type II error (saying it is cavitation when is not) is 3%. By shifting the decision point toward the centroid of non-cavitation (the 'o's), we can improve the performance, i.e., Type I errors can be reduced to 4%, while Type II error will be increased to 10%, if the decision point is set at zero. Realizing the Type I errors here are the critical issue, a more conservative hypothesis could be used. For example, to insure that all cavitation states are detected, a 18% probability would exist for a Type II failure. That is, we would be sure to detect all cavitation (based upon this population), but would be announcing cavitation incorrectly for 18% of the non cavitation states. Our assumption at this point is that these cases (the 18% miss-classified) represent operating conditions which are "near" cavitation and as a result are not desirable. Such assumptions need further study to verify if they are indeed correct.

Table 3: Linear Discriminant Coefficients

P_{in}	-0.02789
P_{out}	-0.00026
T_{in}	-1.60241
T_{out}	1.65222
Q	-0.00247

Conclusions and Comments

1. The SHARP system has successfully demonstrated the ability to do *on-line*, *realtime* system health monitoring of a hydraulic pump.
2. Cavitation for a swashblock piston pump has been characterized, and a technique which relies on the use of a linear discriminant analysis to detect cavitation have been presented.
3. Trending studies would be useful to improve the sensitivity of this detection scheme.
4. Sensor robustness could be the weakest link in the system, sensor failure would lead to a system with no detection capability and as a result the goal of on-line monitoring would be defeated. A study of sensor reliability is needed to understand the limitations of the whole system.
5. The portability of this technology from one piece of equipment to another remains an issue. How specific are the performance characteristic of each pump? How much work is necessary to adapt our present findings to other equipment and other environments, such that SHARP can be installed on other equipment.
6. Future work will look at a comparison of dynamic measurements, i.e., vibration monitoring, with quasistatic system parameters, i.e., pressure, temperature, and flow rate to develop an understanding of their cross correlation as well as ascertaining the potential for sensor fusion of these different measurement..

Acknowledgments

This work could not have been accomplished without the dedicated efforts of many people. I would like to gratefully acknowledge the efforts of all of the following: Bob King, who first recognized the value of System Health Assessment and developed the Line Segment trending representation. Tye Hooley, who did the first linear discriminant

work. Galen Brown and Tim Seifert who ran the experiments and got the data, thanks guys. Morgan Smiley who helped write the system documentation and developed the automated valve control. Rex Rideout, our ace technician who instrumented the test bench and is a handy man to have around. Mark Muzal who built the instrumentation and filtering circuitry and solved the noise problems. Michelle Archuleta whose knowledge of G2 and system diagnosis aided us all. Lori Pratt, who so generously shared her workstation for the initial statistical analysis. Wayne Austad, INEL systems engineer, who's knowledge, informational support, and advice aided our efforts significantly. Mike Donaldson, INEL project manager, who kept us on track, John Richardson, INEL systems engineer, who organized this effort. Kevin Kostelnik, INEL manager, who understands the value in SHA and made this project happen. Finally, I gratefully acknowledge the financial support of Idaho National Engineering Labs, and Sandia National Labs for support of this project.

References

Arnold, K., 1993, "Simple Preventive Checks Can Help Detect and Eliminate Pump Cavitation," *Pulp & Paper*, p.112-114, Feb.

Ding, G., and Hu, D., 1987, "Monitoring and Diagnosis of Typical Faults of Hydraulic Systems," *Condition Monitoring '87*, ed. M.H. Jones, Swansea, U.K., Pineridege Press, p.593-603.

Hooley T. L., 1995, "An Automatic Health Assessment System Prototype," Laboratory for Intelligent Automated Systems technical report, Colorado School of Mines, Golden, Colorado. *LIAS-TR-950301*, March.

King, R. H., and Hooley, T. L., 1993, "System Health Monitoring and Prediction (SHMP) for Telerobotic Equipment Used in Hazardous Environments," Laboratory for Intelligent Automated Systems *technical report*, Colorado School of Mines, Golden, Colorado.

King, R. H., Lever, P. J. A., and Hooley, T. L., 1995, "Reducing Large Volumes of Sensor Data with Linear Approximation," Laboratory for Intelligent Automated Systems *technical report*, Colorado School of Mines, Golden, Colorado.

Mackay, R.C., "Troubleshooting Pump Cavitation," 1993, *Pulp & Paper Canada*, 94:9, p8-10.

Muzal, M. 1996, An Investigation into Noise Suppression of Three Phase Motor Drivers", Laboratory for Intelligent Automated Systems *technical report*, Colorado School of Mines, LIAS-TR-960401, Golden, Colorado.

Pipe, K., 1987, "Application of Advanced Pattern Recognition Techniques in Machinery Failure Prognosis For Turbomachinery," *Condition Monitoring '87*, ed. M.H. Jones, Swansea, U.K., Pineridege Press, p.73-89.

Rubano, T., 1995, "Development of a Real-Time Data Collection System using VMEBus with Ethernet Connections to a Remote Host," Laboratory for Intelligent Automated Systems technical report, Colorado School of Mines, Golden, Colorado, *LIAS-TR-950630*, June.

Smiley, A.M., 1995, "Analysis of the Statistical Character of Hydraulic Pump Parameters from Real-Time Monitoring," Laboratory for Intelligent Automated Systems *technical report*, Colorado School of Mines, LIAS-TR-951201, Golden, Colorado.

Wolf, K., and Ferrara, R.J., 1994, "Investigation of Cavitation Erosion of Different Alloys in Fluids as a Basis for Condition Monitoring of Hydraulic Pumps," *Condition Monitoring '94*, ed. M.H. Jones, Swansea, U.K., Pineridege Press, p.734-747.

QUANTITATIVE DETERMINATION OF THE THICKNESS VARIATION IN THIN SHELL STRUCTURES AND SENSING OF FAILURE POINTS

Iouri Onichtchenko, George Dovgalenko, Anatoli Kniazkov, Greg Salamo
226 Physics Bld, University of Arkansas, Fayetteville, AR 72701

Abstract: In this paper we present a holographic technique which is used for the monitoring of thin cylindrical shell structures. The technique provides a quantitative measurement of the thickness of the thin wall structure and the variation in the thickness of structure. It also provides monitoring in changes in the thickness of the thin shell structure. Failure of the structure is shown to be correlated to the thinnest point in the structure. The holographic technique is based on taking and comparing two holograms of the structure each under different load conditions. For example, if structure is a tank holding corrosive material a hologram is first formed of the tank. A second hologram is than formed with the tank under increased (decreased) load. Comparison between the two hologram than leads to the quantitative measurement of the shell wall thickness as a function of position. The holographic system is based on a thermoplastic holographic camera and a simple HeNe laser. The system is portable and is attached to the structure during measurement. For the case of a cylindrical tank measurement of the wall thickness using the holographic technique comparison a conventional mechanical method agree to within 10%. Our data confirms that failure of the structure occurs at the location at which the wall structure is the thinnest. As a result the instrument can be used to monitor wall thickness and identify expected failure points in thin wall cylindrical, conical, spherical and another structures.

Key Words: Holographic interferometry; shells structure; diagnostics; quantitative measurement; displacements; stresses; stress state

INTRODUCTION: The methods of holographic interferometry present wide opportunities for carrying out non-destructive testing of the quality of various constructions. There are some relatively easy routine applications in large corporation to control tires, composite helicopter blades, components of aircraft engines and to study the vibration of structures to reduce noises [1,2,3,4]. The purpose of such holographic research consists in the detection of the location of defects in the material of construction. Usually in holographic nondestructive testing of real constructions the interpretation of interference patterns has a qualitative nature which consists of analysis of local features of interference pictures as the quantitative measurements of displacements in the real constructions have a large margin of error as the result of uncontrollable motion. The technique eliminates those problems and allows measurement of real constructions in field conditions. The results of the researches from published work on sample lindrical constructions are introduced below.

MECHANICAL PART: If know by Timoshenko [7], solutions of the theory elasticity tasks, where the volumetric forces are constant, is leading et integration of differentiation equation as:

$$\left(\frac{\partial^2}{\partial r^2}+\frac{1}{r}\cdot\frac{\partial}{\partial r}+\frac{1}{r^2}\cdot\frac{\partial^2}{\partial\theta^2}\right)\left(\frac{\partial^2\varphi}{\partial r^2}+\frac{1}{r}\cdot\frac{\partial\varphi}{\partial r}+\frac{1}{r^2}\cdot\frac{\partial^2\varphi}{\partial\theta^2}\right)=0 \qquad (1)$$

for: $\sigma_r=\frac{1}{r}\cdot\frac{\partial\varphi}{\partial r}+\frac{1}{r^2}\cdot\frac{\partial^2\varphi}{\partial\theta^2}$, $\quad\sigma_\theta=\frac{\partial^2\varphi}{\partial r^2}$, $\quad\tau_{r\theta}=\frac{1}{r^2}\cdot\frac{\partial\varphi}{\partial\theta}-\frac{1}{r}\cdot\frac{\partial^2\varphi}{\partial r\partial\theta}$ (2)

where: $\sigma_r,\sigma_\theta,\tau_{r\theta}$ - radial, tangential and shift's stresses respectively,
φ - function of stresses.
Polar symmetric distribution of stresses $(\varphi=\varphi(r))$ is leading from eq (1)

$$\frac{\partial^4\varphi}{\partial r^4}+\frac{2}{r}\cdot\frac{\partial^3\varphi}{\partial r^3}-\frac{1}{r^2}\cdot\frac{\partial^2\varphi}{\partial r^2}+\frac{1}{r^3}\cdot\frac{\partial\varphi}{\partial r}=0 \qquad (3)$$

The particular solutions eq (3) can described as:

$$\varphi=A\ln r+Br^2\ln r+Cr^2+D \qquad (4)$$

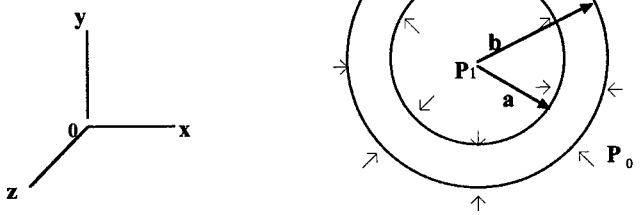


Fig.1. Cylindrical tank with equable load
From expression (4) for boundary condition (Fig.1) : $\sigma_{r=a}=-P_1;\sigma_{r=b}=-P_0$ can be described as the stress-strain state of cylindrical structure with internal load as:

$$\sigma_r=\frac{P_0b^2(r^2-a^2)}{r^2(a^2-b^2)}+\frac{P_1a^2(b^2-r^2)}{r^2(a^2-b^2)} \qquad (5)$$

$$\sigma_\theta=\frac{P_0b^2(r^2+a^2)}{r^2(a^2-b^2)}-\frac{P_1a^2(b^2+r^2)}{r^2(a^2-b^2)} \qquad (6)$$

Where: a,b - inside and outside radius respectively;
r - variable parameter ($a<r<b$); P_1,P_0 - inside and outside pressure.
Radial displacement **w** can be obtained from equation plane stress state [7] as:

$$\varepsilon_\theta = \frac{w}{r}; E\varepsilon_\theta = \sigma_\theta - \upsilon\sigma_r \qquad (7)$$

and it is described as:

$$w = \frac{2P}{E} \cdot \frac{b(b-h)^2}{h(2b-h)} \qquad (8)$$

Where: h = b - a

HOLOGRAPHIC PART: Holographic interferometry provides the means to detect the displacement produced by the stresses in materials. The reader is refereed to example [1], [5]. If the hologram of the surface is made under required load, then the hologram is exposed again or viewing under real-time conditions, the result is a pattern like shown in Fig.4. Each light or dark fringe represents contours of constant phase shifts. At a given point in material, the relationship between the phase shift and displacement is given by [6]

$$\vec{K}\Delta\vec{r} = k\lambda \qquad (9)$$

where: k - number of interference fringe; $\vec{K}$ - sensitivity vector; λ - wavelength; $\Delta\vec{r}$ - displacement vector.

The sensitivity vector $\vec{K}$ is dependent on the illumination and viewing of the parameters of the experiment and is found as the vector difference of the respective propagation vector:

$$\vec{K} = \vec{\rho}_2 - \vec{\rho}_1 \qquad (10)$$

where: ρ_1, ρ_2, - the propagation vectors defining the directions of the illumination end viewing respectively (Fig. 2).

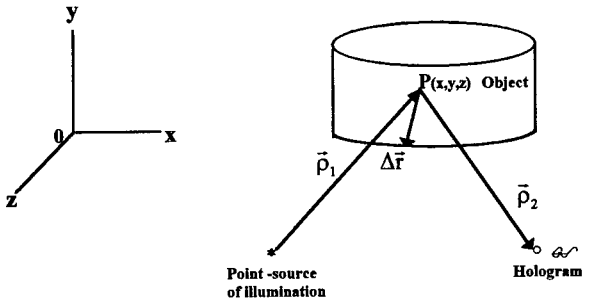


Fig.2. Illumination and observation geometry

Equation (9) can be defined in respect to the origin of the Cartesian X-Y-Z coordinate system as:

$$Mu + Nv + Lw = \lambda k \qquad (11)$$

Where: $M = \cos\varphi_0 \sin\theta_0 - \cos\varphi_1 \sin\theta_1; N = \sin\varphi_0 \sin\theta_0 - \sin\varphi_1 \sin\theta_1;$

$L = \cos\theta_0 - \cos\theta_1$

or:

$$a_{11}u + a_{12}v + a_{13}w = k\lambda \qquad (12)$$

Where: θ_1, θ_2 - angels between OZ-axis and of illumination and observation orts respectively; φ_0, φ_1 - angels of the OX- axis to the projections of the illumination and observation orts on the XOY- coordinate plane; u, v, w - projections of displacement vector $\Delta\vec{r}$ on Cartesian X-Y-Z coordinate system respectively; a_{11}, a_{12}, a_{13} - coefficients of optical schema are determined as:

$$a_{11} = \frac{x - x_0}{\sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}} - \frac{x_1 - x}{\sqrt{(x-x_1)^2 + (y-y_1)^2 + (z-z_1)^2}}$$

$$a_{12} = \frac{y - y_0}{\sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}} - \frac{y_1 - y}{\sqrt{(x-x_1)^2 + (y-y_1)^2 + (z-z_1)^2}} \qquad (13)$$

$$a_{13} = \frac{z - z_0}{\sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}} - \frac{z_1 - z}{\sqrt{(x-x_1)^2 + (y-y_1)^2 + (z-z_1)^2}}$$

Thereby in order to completely determine its unknown components u,v,w from equations (12) an "ideal" system of three equations of the type (12) must be solved simultaneously. Such equation can be generated using the observation for various directions. For the case of determination of displacement on the surface of cylindrical shell, the displacements are functions of load and thickness:

$$u = f_1(P,h); \; v = f_2(P,h); \; w = f_3(P,h) \qquad (14)$$

where: P- internal pressure; h- thickness of shell.
Using (14) equation (12) can be expressed as:

$$a_{11}\, f_1(P,h) + a_{12}\, f_2(P,h) + a_{13}\, f_3(P,h) = k\lambda \qquad (15)$$

The equation (15) is describing the depends of interference patterns of the tested area from internal pressure and thickness of construction. From eq (15) quantity **h** can be determined if P=const., or quantity P $\langle \sigma_x(P), \sigma_y(P) \rangle$ if h = const for each point of surface from one direction of illumination and observation.

DISPLACEMENT RELIEVED FROM THE DIFFERENT THICKNESS OF CYLINDRICAL HOLDING TANK WITH EQUABLE (HYDROSTATIC LOAD: If $P_0 = 0$ and the cylindrical shell has only internal load expression (5) and (6) for b = r (on the surface of the cylindrical shell) can be described as:

$$\sigma_r = 0; \sigma_\theta = \frac{2a^2 P_1}{b^2 - a^2} \qquad (16)$$

In the case, when cylindrical shell was shut and consistent under action of the internal pressure displacement was expressed as:

$$w = w_1 - w_2 = \frac{2P}{E} \cdot \frac{b(b-h)^2}{h(2b-h)} - \frac{\nu P}{2E} \cdot \frac{(2b-h)^2}{4h} \qquad (17)$$

Where: w_1 - is obtained from equation (8); $w_2 = -\frac{\upsilon P}{2E} \cdot \frac{(2b-h)^2}{4h}$ – displacement from the load aboard cylindrical axis.

In the case of maximum sensitivity of the holographic interferometer to out-plane displacements equation (15) considering (11) can be described as:

$$a_{13} \cdot \frac{P}{E}\left[\frac{2b(b-h)^2}{h(2b-h)} - \frac{\upsilon(2b-h)^2}{8h}\right] = k\lambda \qquad (18)$$

For inside pressure (axis-semetric load) some points of surface having equal displacement are concurrent. Difference between displacement of those points is equal. For example: for cylindrical tanks there are points located on the circumferences of cross sections normal to the cylindrical axis. Considering those conditions from equation (15), can be described:

$$a_{13}^1 w_0^1 - a_{13}^0 w_0 = \Delta k_0 \lambda \qquad (19)$$

$$a_{13}^1 w - a_{13}^0 w_0 = \Delta k \lambda \qquad (20)$$

Subtraction eq (19) from eq (20) yields:

$$\Delta w = \frac{\Delta k - \Delta k_0}{a_{13}^1} \lambda \qquad (21)$$

where: $\Delta w = w - w_0^1$; a_{13}^1, a_{13}^0 - coefficient researched points in/out zone of the feature

interference pattern; w_0 - displacement of points are locating out zone of the feature interference pattern; w, w_0^1 - displacement of points with/without availability of variations of thickness respectively; $\Delta k_0, \Delta k$ - relative order of interference respectively; λ - wavelength of the source.
Substitution eq. (17) in deduction eqs (19) and (20) is giving equation of third power:

$$h^3 + Bh^2 + Ch + D = 0 \qquad (22)$$

Where: $B = \frac{1}{\upsilon}(b^2 - 6b\upsilon + 2H)$; $C = \frac{b}{\upsilon}(-2b + 12\upsilon b - 4H)$; $D = b^3(\frac{2-\upsilon}{\upsilon})$;

$$H = \frac{\lambda E(\Delta k - \Delta k_0)}{Pa_{13}^1} + \frac{2b(b-h_0)^2}{h_0(2b-h_0)} - \frac{\upsilon(2b-h_0)^2}{2h_0}$$

From solution eq (22) can get quantity volume of the thickness of the separate site of of the shell in the case of closed cylindrical construction under efficient inside load. In the case of "opened" cylindrical constructions under efficient inside hydrostatic load the expression (22) can be described as equation of second power:

$$h^2 + ph + q = 0 \qquad (23)$$

Where: $p = -2b$; $q = \frac{2Pb^3h_0(2b-h_0)a_{13}^1}{\lambda(\Delta k - \Delta k_0)Eh_0(2b-h_0) + 2Pb^3a_{13}^1}$.

The radical of eq (22) can sense for it case describe as:

$$h = b - \sqrt{b^2 - \frac{2Pb^3h_0(2b-h_0)a_{13}^1}{\lambda(\Delta k - \Delta k_0)Eh_0(2b-h_0) + 2Pb^3a_{13}^1}} \qquad (24)$$

Where: P - variation of the load for record of the objects interferogram.

IIIUSTRATIVE EXAMPLE: In order to verify the ability of the holographic technique to determine of the vessel's thickness, a series of experiments were conducted using real cylindrical construction, and for each case the thickness of the wall was decreased in deferent locations. The large number of the cylindrical constructions were monitored, for vessels of different materials (steel, aluminum), geometrical parameters, and wall thickness. The aim of the research was the following:
a) determination of the location of zones with different thickness; b) determination of the thickness in each zone through holographic technique; c) comparison of holographical and mechanical methods of measurements. The results of only one of the experiments will be considered here for the purpose of illustrating the application of the theoretical approach derived in this paper. In Fig.3 is the steel cylindrical vessel which was monitored ($E = 30.0 \times 10^6$ psi; $\nu = 0,3$; $b = 3.0$in; $h_0 = 0.012$in)

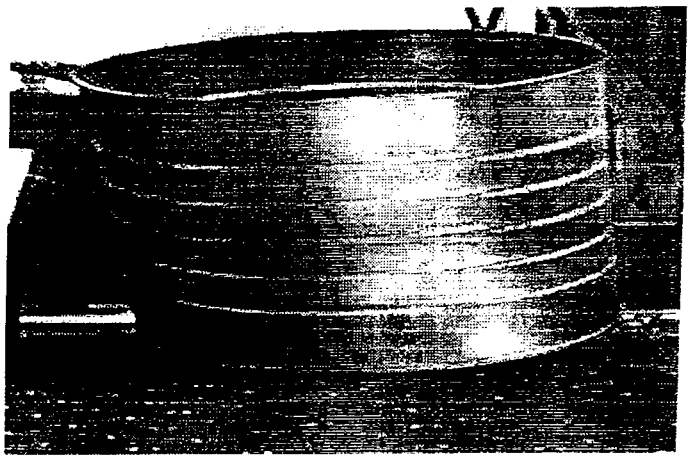


Fig.3. The photopicture of the steel's cylindrical vessel

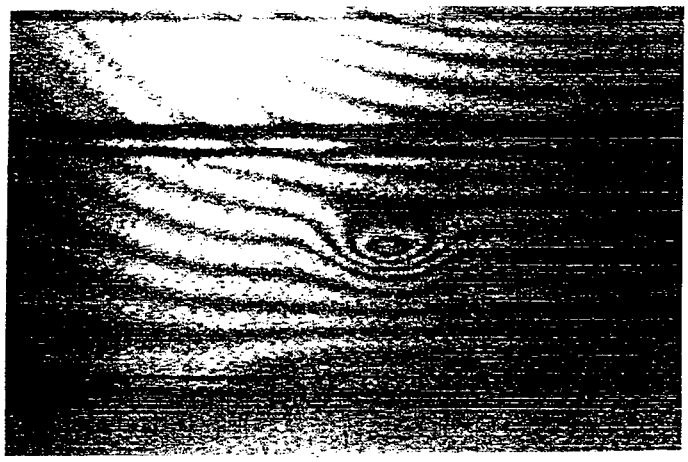


Fig.4. Interferogram visualisation decreased thickness zon in the wall of the cylindrical vessel

The experimental fringe pattern of detected thickness of the zone in the wall is presented on Fig.4. In this case we applied the internal water loading. As a result we see interference fringes between unloaded and loaded state of monitored object. The results of measurement decreased thickness:
(h) = 0.002954in - holographical measurement;
(h) = 0.003000in - mechanical measurement.

CONCLUSION: Experimental application of described holographic technique shows that it allows for monetoring real constructions. In our experiments inaccuracy of holographic method was < 10%. Noncontact holographic method is a large advantage of measurement.

References

1. Vest, C. Holographic Interferometry, J.Wiley, New York (1979).
2. Ostrovsky, Yu., Shchepinov, V.P., and Yakovlev, V.V. Holographik Interferometry in Experimental Mechanics, Springer Verlag, Berlin (1991).
3. Ennos, A.E. Holographic Nondestructive Testing with EW Lasers Away from the Laboratory. Holographic Nondestructive Testing (1992).
4. Sciamarella Cesar A., Naraynan K. Holographic Techniques for Detection of Flawsin ceramics. Proc. S.Int.Cong. Exp. Mech. Mountreal. Jun.10,!984,p.17-20.
5. Pryputniewicz R.J. Quantitative Deterrmination of Displacement and Strains from Holograms. Springer Series in Optical Sciences, Vol.68,R.K.Rastogy (Ed.), Holographic Interferometry, Springer- Verrlag Berlin Heidelberg !994.
6. Aleksandrov E.B., Bonch- Bruevich A.M.: Sov. Phis.Tech. Phys.12,p.258-265,(1967)
7. Timoshenko S.P.,Goodier J,N. Theory of elasticity. Naw -York, McGram-Hill,(1969).
8. Birger I.A. The sircular's plates and shells of the rotation. M., Oborongiz, 1964.

AN AUTONOMOUS DIAGNOSTIC/PROGNOSTIC SYSTEM FOR SHIPBOARD CHILLED WATER PLANTS

Thomas G. Edwards and George D. Hadden, Ph.D.
Honeywell Technology Center
3660 Technology Drive
Minneapolis, MN 55418

Abstract: This paper describes an autonomous system for diagnosing faults and predicting the remaining life of shipboard chilled water plant components. The goals of this system are to provide increased machinery diagnostic/prognostic capability at a time when available maintenance manpower is being reduced. The system will provide increased mission readiness and support maintenance scheduling by providing an accurate, autonomous assessment of machinery condition. This open diagnostic/prognostic system does not focus on a single condition based maintenance technology but is designed to combine multiple, diagnostic/prognostic approaches that will, individually, provide different information about the machinery's condition. This system is designed to blend multiple approaches and will also accept new technologies and algorithms as they are developed or become available. Sensor and sensor validation algorithms will be designed and chosen that provide reliability, redundancy and low cost. Local processors will be used to autonomously and continuously collect data and monitor the chilled water plant components. A centralized processing and data repository will also provide prognostic results to maintenance personnel. This paper will discuss the three sub-systems of the diagnostic/prognostic system: sensors, Data Concentrators (DCs) and the Prognostics/Diagnostic Monitoring Engine (PDME).

Key Words: Chilled water plant; diagnostics; knowledge fusion; MEMS; prognostics; sensor fusion; vibration analysis

INTRODUCTION Honeywell, in partnership with the Office of Naval Research (ONR), The Naval Research Laboratory (NRL), The Naval Surface Warfare Center (NSWC), PredictDLI, AlphaTech, Knowledge Partners of Minnesota, the Center for Integrated Diagnostics, and WM Engineering Ltd, is developing a distributed shipboard system to perform diagnostics and prognostics on rotating equipment (e.g. engines, purifiers, pumps, generators, chillers). The goal is to significantly impact the economics and efficiency of maintenance operations and logistics. Achieving this goal enables an increase in the overall readiness of the Navy's fleet, and at the same time allows operational man-power requirements to be reduced. This Condition Based Maintenance (CBM) system, called MPROS for *Machinery Prognostics/Diagnostics System,* consists of MEMS and conventional sensors on the machinery, local intelligent signal

processing devices, called Data Concentrators (DCs), and a centrally located subsystem called the PDME, for *Prognostics, Diagnostics, Monitoring Engine.* MPROS augments periodical vibration analysis by collecting data continuously from vibration and other sensors, These data streams are integrated as necessary in the Data Concentrators with data fusion techniques. Knowledge fusion, a second level of integration, occurs in the PDME. At this level, the outputs of different diagnostic and prognostic reasoning mechanisms are fused to yield the best possible analysis. In this paper we discuss the algorithms and hardware required to support the system

The objective of the Honeywell program is to develop an automated diagnostics/prognostics system that provides Navy personnel with a reliable and accurate means of making maintenance decisions based on the physical condition of components in shipboard machines. A suite of sensors, some based on MEMS (micro electro-mechanical systems) technology, monitor the behavior of each component and its immediate environment. Signals from these sensors are brought together in a Data Concentrator (DC), mounted near the machine, that interprets the signals and applies algorithms to the data to perform data reduction, diagnostics, and limited prognostics, and local display of machinery condition. The outputs of the DCs are transmitted to a prognostics computer which performs the necessary computations to analyze the many sets of reduced data, and determine a remaining life value for all of the machinery components being monitored. The prognostics for each component relies on the optimal combination of heuristic and physical models of the failure process.

TECHNICAL APPROACH In this paper we define diagnosis as fault recognition and identification (i.e., the processes used to recognize that a symptom is indicative of an underlying machine fault and then determining what fault is causing the symptom). Prognosis, on the other hand, is the prediction of failure and is similar to diagnosis, but there are three differences: In diagnosis, there is a starting point (i.e., a symptom of the fault that triggers the activity). Prognosis is not triggered but rather is performed continuously or periodically for all equipment of interest. Prognosis supplies an estimate of when the device will fail. To predict failure, parameter values that are considered normal must be examined more closely for subtle variations and for relationships with other parameter values. Assessment of "normal values" becomes more constrained.

Figure 1 shows a block diagram of the MPROS system, which consists of three physically separate subsystems: A suite of sensors, some of them MEMS-based, The Data Concentrators (DCs) that do intermediate processing local to the machinery, and The prognostic/diagnostic/monitoring engine (PDME). Two of these subsystems, the PDME and the DC, contain software. The sensors themselves supply digital or analog output that becomes the input to the DC.

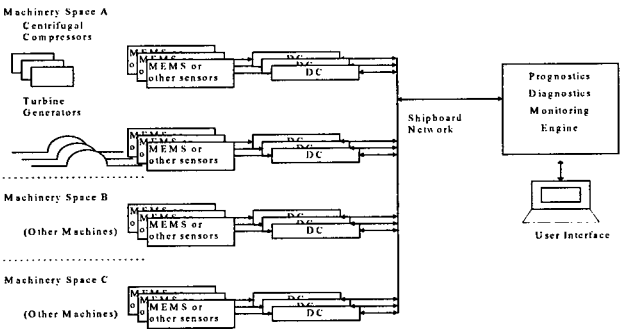


Figure 1. MPROS System Architecture

SENSORS MRPOS uses a mixture of existing and new sensors and sensing techniques to achieve the best prognostic/diagnostic results. The required sensor suite varies for different machinery types and configurations. The approach includes the use of proven existing sensor and signal processing techniques to guarantee a viable product, the development of an integrated MEMS system and the blending of multiple sensors via sensor fusion to provide a more complete answer than possible with a single sensing device.

Sensors must be selected that most effectively interrogate the components for faults. The machinery must be instrumented with the optimal type, number, and mixture of new and existing sensors and sensing techniques. Our sensor approach is based on use of Commercial off the Shelf (COTS) sensors, as well as the development of new sensors for fault detection to meet the needs of the CBM system that cannot be met by conventional COTS sensors, using only those sensors that are necessary and sufficient, and using existing, installed sensors where possible. Parameters to be measured include machinery performance, physical positions of equipment, ambient environmental input factors, machinery generated parameters, fugitive emissions, structural failure mechanisms, interaction between machines, inputs to the machine such as fuel, water, lubricants, contaminants, and electrical power quality, and corrosion and fouling factors.

THE DATA CONCENTRATORS Data Concentrators (DCs) are small, rugged devices that are installed near the machinery. They take as input raw sensor data, digital sensor data, or output from other DCs. Software resident on the DC includes State-Based Feature Recognition and neural network software for local diagnostics and prognostics, signal processing (e.g., wavelets and Fast Fourier Transforms (FFTs)), and data reduction. The output of the DCs is transmitted to the PDME and is available locally to provide an immediate warning indicator. The DC's software can be

downloaded from the PDME, so that the behavior of its algorithms can adapt based on knowledge gained from other machinery, or so that data-acquisition and processing capabilities can be modified. A block diagram of the DC is shown in Figure 2.

DATA CONCENTRATOR HARDWARE ARCHITECTURE The design of the acquisition system is based on the sensory inputs requirements for the specific machinery. A single DC is capable of simultaneously interfacing to multiple sensors and multiple sensor types. In addition a DC can monitor data parameters on more than one machine, or, alternatively, they can be "cascaded", i.e., one DC may monitor the outputs of two or more other DCs). This capability is required when local processing requires timely delivery of data from more than one machine. The alternative of routing raw data to PDME and back to one of the DCs is supported inthesystem but usually is too cumbersome.

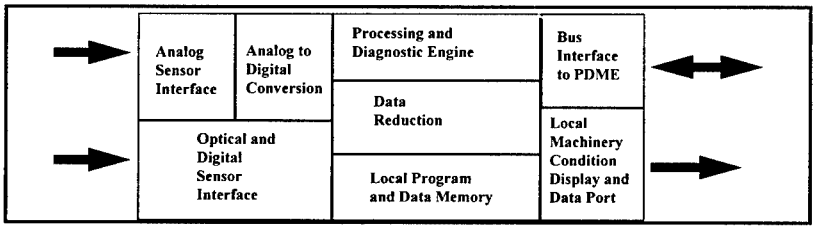


Figure 2. Block diagram of data concentrator

Acquisition requirements for a single DC may range from low-bandwidth, high-resolution strain gauge inputs to high-bandwidth acoustic emission measurements.
The DC consists of three subsystems: Data acquisition system (including the sensor interface and digitization stages) Diagnostic processor (consisting of a DSP or microcontroller) Bus interface (comprising an interface to a portable debrief device such as a PC or Personal Digital Assistant (PDA) and an interface to electrical or optical data bus) Continuous acquisition and processing provide a series of signatures of the machinery's data parameters over the operational conditions of the equipment.

Data are not retrieved at single-point instances in the machines life. The continuous monitoring approach provides a life-time environmental and trend analysis of the machinery to the prognostic system. DCs provide a path for total autonomy of the diagnostic/prognostic system, requiring no human involvement to manually collect and transport data. Sensor data bandwidth is decreased by orders of magnitude, decreasing prognostic algorithm and data transfer requirements. Raw, compressed, and processed

data are stored in non-volatile memory local to the DC. The selection of the type and size of the local memory device is based on prognostic/diagnostic requirements.

An additional important autonomy exhibited by the DC is the characteristic that if PDME the DC continues to perform in a robust way. In particular, local alerting is not interrupted. The DC hosts both diagnostic and statistical algorithms described below. It can dynamically receive and execute new algorithms downloaded from the PDME, as required. The DC includes two features to ensure that an accurate assessment of the equipment health is provided at an affordable level: it is capable of performing a complete diagnostic survey (built-in test or self-diagnostics) of itself, and machinery condition can be accessed through a local data port on the DC. This port can also be used in conjunction with a hand-held device, similar to a personal digital assistant.

DC Software Architecture: The DC can host a set of algorithms consisting of data reduction and storage, signal processing (including both time-frequency and normalized vibration analysis), pattern recognition, data trending, and a limited form of prognostics and diagnostics. In addition, becausethealgorithms can perform more effectively with several coordinated sensor inputs rather than a single one, we have included the capability to *fuse* sensor inputs. Some of these algorithms are described here:

Data Reduction and Storage:. Data is available to both the DC and the PDME for prognostic/diagnostic processing in both raw and reduced formats. Historical and trended data parameter values are maintained in the DC for retrieval by the PDME. Data is reduced and stored using a variety of compression and time-sampled methods including predictors, histograms, and circular buffers.

Signal Processing Techniques: A variety of signal processing techniques are employed in the DC to provide inputs to the prognostic/diagnostic algorithms. Among these are wavelet transforms and FFTs. In general, these kinds of techniques are most useful for parameters with a high rate of change (e.g., vibration and acoustic emission signals). Lower frequency signals (many of which are constant) generally do not require signal processing techniques, but are instead fed directly into the prognostic and diagnostic algorithms.

Sensor Fusion: Two primary sensor fusion methods are included in the architecture: SBFR and Hierarchical Neural Networks. Although both of these techniques take input data from multiple sensors, their capabilities extend well beyond sensor fusion. Both perform local diagnosis and prognosis and have the capability of informing the SMDE of anomalous behavior. They are described below.

State-Based Feature Recognition (SBFR): SBFR is a Honeywell-developed technique for the hierarchical recognition of temporally correlated features in multi-channel input. It consists of a set of several enhanced finite-state machines operating in parallel. Each state machine can transition based on sensor input, its own state, the state of another state machine, measured elapsed time, or any logical combination of these. This implies that systems based on SBFR can be built with a layered architecture, so that it is possible use them to draw complex conclusions, such as prognostic or diagnostic decisions.

The implementation of the SBFR system is ideal for embedding into the DC. We have successfully applied SBFR-based diagnostic and prognostic modules to several problems and platforms, including valve degradation and failure prediction in the Space Shuttle's Orbital Maneuvering System, imminent seize-up in Electro-Mechanical Actuators through electrical current signature analysis and other parameters, and failure prediction in several subsystems (including Control Moment Gyro bearing failure) in the Space Station Freedom's Attitude Determination and Control System.

SBFR embedded in the DC takes as input the raw sensor data and the output of other algorithms (e.g., DLI's vibration analysis and FFTs) and performs trending analysis, feature extraction, and some diagnostics and prognostics on this data. Local knowledge gained by these techniques is delivered to for further analysis and crew alerting. Local alerting (e.g., indicator lights) is also available. Under control of the System Executive running in the PDME, new finite-state machines may be downloaded into the smart sensor. This allows the behavior of the sensor to adapt to its data in appropriate ways.

Hierarchical Neural Networks: Because of the nonlinear transfer function inherent in many machinery applications between vibration source point and the sensor location, which is usually located on the machinery housing, multiple sensors are often required to provide complete fault coverage. One natural way to take advantage of these multiple sensors is to improve the overall FDI (fault detection and isolation) through the use of hierarchical neural networks. Design of an appropriate hierarchical neural network to learn an appropriate decision-making strategy produces a more robust and overall higher performing FDI system when three or more sensor sites are used. This fusion strategy has been applied to space shuttle main engine high-pressure oxygen turbo pumps, CH-46 gearbox, Sikorsky S-61 and S-71 main transmissions. The results have consistently shown that implementing this type of sensor fusion results in higher performing systems. Clearly the outputs of such systems can be aggregated over time to achieve very high performing systems with extremely low levels of false alarms.

Other Prognostic and Diagnostic Algorithms: This list of potential DC-resident fusion software is not exhaustive. Thearchitecture allows for the inclusion of new and different techniques as they are required.

THE PDME: The prognostic/diagnostic/monitoring engine (PDME) is a set of cooperating software modules. It can collect data, apply prognostics and diagnostics algorithms to the data, and provide its analyses, including warnings, incipient failure recognition, and machinery condition. Sensor and knowledge fusion are used to extract the best possible information from the data supplied. The PDME modules are shown in Figure 3 and described in detail in the following subsections. They include:

The Object-Oriented Ship Model: The object-oriented ship model (OOSM) represents the ship's components and a number of relationships among them. It serves as the communications mechanism among the other modules and provides ship data to them.

Knowledge Fusion: Honeywell has built a diagnostic/prognostic system that uses Dempster-Shafer rules of evidence to combine knowledge from a number of concurrently executing diagnostic and prognostic algorithms. We are using a similar system in the PDME to combine the outputs of other modules and to provide a pathway by which new modules can be inserted. Together with the SE and the OOSM, this system allows new methods to be added in a systematic way.

Vibration-Based Prognostics Algorithms: Both a rule-based system and a neural network approach reside in the PDME, the former for steady-state analysis and the latter for transient analysis.

State-Based Feature Recognition: SBFR is used both in the DCs and in the PDME. In the PDME, it combines data from individual machine sensors, the analysis performed by DCs, and ambient environmental conditions to draw diagnostic and prognostic conclusions.

Historical Analysis: As the system runs, it automatically collects data relating symptoms to faults and augments existing fault models. This data is stored in the OOSM and used in diagnosis and prognosis.

The System Executive: The SE coordinates the activities of the other modules as well as administers the knowledge fusion activity and the OOSM.

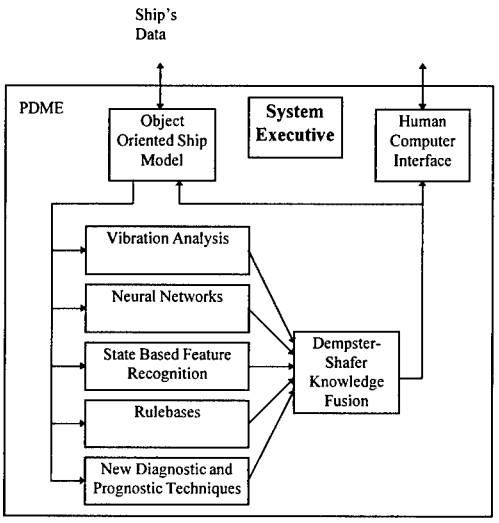


Figure 3. PDME Block Diagram

A COMPLEMENTARY SUITE OF DIAGNOSTIC AND PROGNOSTIC ALGORITHMS Shipboard machinery diagnostics/prognostics is a domain of considerable complexity. A variety of machine types must be considered: availability of fault data and fault models varies considerably, different sensor suites need to be handled, criticality of fault-free operation is a function of the machine and of the nature of the ship's operation, and so forth. The complexity of the domain argues strongly against single-technology solutions. No one approach can satisfactorily address the full variety of the problems and constraints.

MPROS is a system incorporating several key diagnostic/prognostic approaches. These approaches have been selected to individually provide powerful techniques for handling specific maintenance problems and to collectively create a synergistic integration of methods that, in combination, address the full complexity of the shipboard condition-based maintenance domain.

MACHINERY FAULT MODEL DEVELOPMENT Machinery fault models are based both on theoretical and empirical investigation. Theoretical models are derived from mathematical system models and lead to useful and accurate broad fault data definitions. For example, the theoretical effect of misalignment across a mechanical

coupling can be modeled accurately knowing the stiffness of the coupling, the shafting, bearings, housings, and foundation. Additionally, changes in vibration patterns caused by imbalance are also predictable theoretically as are the vibration patterns exhibited by various faults occurring in rolling contact and journal bearings.

In reality however, mechanical system nonlinearities due to friction, gaps, stiffness discontinuities, and damping all yield inaccuracies in theoretical predictions. This is largely because these qualities are difficult to model and small changes between the model and reality yield significant differences between predicted and actual system behavior. For imbalance, and misalignment, this inaccuracy is negligible for most cases, but it can be a dominant factor in some cases. For bearing problems, both rolling contact and journal, the nonlinearities contribute extensively to the pattern of vibration and yield significant variation in the vibration spectra exhibited as a fault progresses in severity.

Empirical means of fault modeling bridge the gap between theoretical and actual fault indications. Vibration, temperature, and other measurements made prior to machinery breakdown and subsequent machinery repair feedback are an invaluable source of mechanical fault modeling. In these cases, the fault model is not a model at all, but is the actual machine with an actual fault. It serves as a model for similar faults as they occur in identical and other similar machines. Theoretical fault modeling demonstrates forms which the data should take when a given fault is present and empirical methods bridge the gap by demonstrating the unpredictable variations from theory which occur in reality.

RSAFP

DESIGN, FAILURE PREVENTION AND STRESS ANALYSIS

Chair: **Shahriar Jahanian**
Temple University

On the Redesign and Material Selection of a Pedal Spindle Used in Bicycles: A Case Study

M. Manoogian, D. Johansen, P. Crawford,
R. Nishimuro, J. Foyos, F. Fisher
and O.S. Es-Said
Loyola Marymount University, Los Angeles, CA 90045

Abstract

A case study, obtained from McMahan Titanium Company in California, examines the failure of a pedal spindle for use on a standard Titanium alloy B265 grade 5. A stainless steel ring is press fit onto the spindle as a support for the bearings. The spindle is designed to hold the bearing required for the pedal itself. A plastic sleeve slides over the spindle to provide a threaded surface for the attachment of the pedal. A stainless steel race with ball bearings sets over the stainless steel ring already on the spindle. The bearing covers connections between the two sets of bearings and is kept in place with a small attachment. The whole assembly is kept in place when the pedal is attached. A problem occurred when the pedal spindle began to fail at the point where the stainless steel ring is attached. This failure was apparently a fast fracture of the solid titanium alloy shaft. The reason for the constant changes in diameter of the original design was to maintain a minimum weight for the spindle. A design change is suggested in this report along with an analysis of the material selection.

Key Words: Modeling, Materials Selection, Ashby's Technique, fatigue analysis, stress concentration factors.

Introduction

This case study examines a bicycle pedal support with a high incidence of failure consistently at the same section. The pedal spindle, machined

from bar stock Ti alloy B265 grade 5, Figures 1(a) and (b) to handle the bearing stresses from the ball bearings. During use, failure occurred at the sudden diameter change in the spindle from 0.340 in. to 0.250 in. No appreciable fillet was machined into the corner created by the diameter change. A stress analysis was done on the section.

STRESS ANALYSIS

Assumptions in this stress analysis include a transverse force of 250 lbs. centered 0.3 in. from the location of fraction, fatigue is the worst case design consideration, stress concentrations affect the strength at the location of fracture, and loads are transmitted by the bearings. As a result, a free body diagram may developed as shown below in Figure 2:

Apply equilibrium equations assuming counter-clockwise moments are positive:

$$\Sigma F_y = 0 = -125\text{lb} - 125\text{lb} + F$$

$$F = 250\text{lb}$$

$$\Sigma M_z = 250\text{lb}(0.3") = 75\text{inch-lb} \qquad (1)$$

Determine the maximum shear stress, τ where:

$$\tau = \frac{F}{A} \qquad (2)$$

where $A = 0.0491$ in^2, the cross-sectional area for the 0.25 inch diameter shaft. The shear stress is then 5092 psi. The bending stress is then:

$$\sigma = \frac{M_z C}{I_{zz}} = \frac{(75\text{inch-lb})(0.125\text{inch})}{\frac{\pi(0.125\text{inch})^4}{4}} \qquad (3)$$

$$\sigma = 48{,}917 \ \frac{\text{lbs}}{\text{inch}^2}$$

where C is the radius of the shaft, I_{zz} is the moment of inertia and σ is the bending stress. The spindle is machined from a Ti alloy with 6% aluminum and 4% vanadium by weight. Yield and ultimate tensile strengths for this material are 140ksi and 150ksi, respectively, [1]. The endurance, S_c, is half the ultimate strength or S_c'=75ksi for 1 million cycles, [2].

To determine the endurance limit, the endurance strength, S_e, is scaled as follows, [2]:

$$S_e=K_aK_bK_cK_dK_eS'_cK_R \qquad (4)$$

where

K_a=Surface Factor

K_b=Size Factor

K_c=Load Factor

K_d=Temperature Factor

K_e=Miscellaneous Factor

K_r=Reliability Factor

Factors K_c, K_d, and K_e where assumed to be 1. The surface of the spindle has a machined finish. For the material, $K_a = aS^bULT$. a = 2.7, b = -0.265, then K_a = 0.716, [2]. The size factor, $K_b = (d/0.3)^{-1133}$, d = 0.25in., then k_b = 1.02, [2]. The reliability factor, K_R, was based on a 99% reliability assumption, then K_R = 0.814, [2]. The adjusted endurance strength after annealing is then:

$$S_e=(.716)(1.02)(1)(1)(1)(.814)(75,000lb/inch^2)$$
$$=44,586lb/inch^2 \qquad (5)$$

The stress concentration exists at the point of fracture (Figure 1b), where a very small fillet radius of about 0.007" exists. The diameter, D, of the shaft at the fracture point (Figure 1a and

1b) is 0.340". The diameter of the stepped down section of the shaft is 0.25". For ratios D/d = 1.36 and r/d = 0.028, the stress concentration factor, K_T, is 2.42 [2]. For the alloy, the notch sensitivity factor, q, for r = 0.007", and an ultimate stress of 150ksi is 0.66. The adjusted stress concentration factor, K_F is then, [2]

$$K_f=1+q(K_T-1)=1+0.66(2.42-1)=1.937 \quad (6)$$

The stress concentration correction factor, K_s, is then:

$$K_s=\frac{1}{K_F}=0.516 \quad (7)$$

Thus the endurance strength, S_c, corrected for the stress concentration is then:

$$S_e=K_sS_c=23,006 lb/inch^2 \quad (8)$$

Based upon the assumption and calculations presented previously, the part is not strong enough for the expected design conditions since the maximum applied stress of 49,418psi exceeds the corrected endurance strength of 23,006psi. In order for the design configuration to be sufficient, the ultimate tensile strength of the material must exceed 322,000psi. Alternatively, a reduction of the stress concentration may be achieved by requiring a larger fillet radius. If this radius is a minimum of 0.025 in, the stress concentration factor becomes 1.64, [2] and the stress concentration reduction factor becomes 0.656. The corrected endurance strength is then:

$$S_e=(0.656)\ (44,586 lb/inch^2)=29,248 \frac{lb}{inch^2} \quad (9)$$

This is still much less than the required strength of the spindle. In addition to the increase in the fillet radius of the spindle, polishing out

the machining marks would remove scratches. As a result, K_a would be 0.875, [2]

$$S_e=(0.875/0.716)\ (29,248\text{lb/inch}^2)=35,743\text{lb/inch}^2 \qquad (10)$$

This is still significantly lower than the required strength of 48,917psi. The remaining options would include the selection of a material with an ultimate strength of (48,917/35,743) (150,000psi) or 205,287psi.

MATERIAL SELECTION

The preliminary material search involved use of Cambridge Material Selector, [3], a computer program that determines which group of materials will fit a design and application given specific performance indices, C_i, i=1,2,3,4,5. The following derived performance indices for a beam using Ashby's Technique, [3,4] were:

$$C_1=\frac{\sigma_y^{2/3}}{\rho};\ \text{Maximizes Strength/unit weight}$$

$$C_2=\frac{E^{1/2}}{\rho};\ \text{Maximizes Stiffness/unit weight}$$

$$C_3=\frac{\sigma_{\delta y}^{2/3}}{\rho C_R};\ \text{Maximizes Strength: Weight Index/cost}$$

$$C_4=\frac{E^{1/2}}{\rho C_R};\ \text{Maximizes Stiffness: Weight Index/cost}$$

$$C_5=\frac{K^{2/3}}{\rho};\ \text{Maximizes Fracture Toughness/Unit Weight} \qquad (11)$$

where σ_y is the yield strength, ρ is the density of the material, E is the modulus of elasticity, C_R is the relative cost, and K is the fracture toughness.

The search was further limited to materials with an ultimate strength greater than 210ksi, (stage 6). Given the performance indices and material

selection charts, see for example Figure 3, it was determined that titanium alloys, carbon fibers in epoxy, or carbon reinforced polymers passed all six stages and would best suit the spindle design. Since only a limited number of pedals would be produced, carbon fiber and carbon reinforced polymers will be disregarded for their high production costs.

A titanium alloy with a suitable strength was found at Titanium Metals Corporation. The TIMET alloy Ti-1.5A1-4.5Fe-6.8Mo can have an ultimate strength higher than 220ksi with the proper heat treatment. The heat treatment consisted of solutionizing the TIMET alloy at ~ 735°C (1355°F) for 10 minutes and aging at ~ 510°C (950F) for eight hours. The ultimate strength, yield strength, total percent elongation and reduction in percent area of the alloy are ~ 222ksi, 215ksi, 3%, and 15% respectively [5].

CONCLUDING REMARKS

The failure of a pedal spindle used in bicycles was analyzed. A slight design modification was suggested to reduce the stress concentration factors. A material with a minimum of 205.2ksi in ultimate tensile strength was sought for. Ashby's technique of material selection was used. Material property performance indices were suggested for a beam that would maximize its specific strength, specific stiffness, specific fracture toughness and with minimum cost. Composites and Titanium alloys passed all stages of the performance indices, however composites were rejected due to their high production costs. Titanium manufactures were contacted and a specific titanium alloy was found, together with the design modifications, to apparently prevent the recurring failure.

REFERENCES

1. D. R. Askeland, The Science and Engineering of Materials, p: 407 (Boston; PWS, 1994).

2. J. Shigley and L. Mitchel, Mechanical Engineering Design, (New York: McGraw Hill Book Company, 1983).

3. Cambridge Materials Selector, version 1.1, Software, W. Chong, R. Thomas, D. Cebon, and

M. Ashby, Cambridge University Engineering Department, Cambridge, U.K. 1990.

4. M. Ashby, Materials Selection in Mechanical Design, (New York: Pergamon Press, 1992).

5. Titanium Metals Corporation, data specifications sheets, 1999 Broadway, Denver, CO 80280.

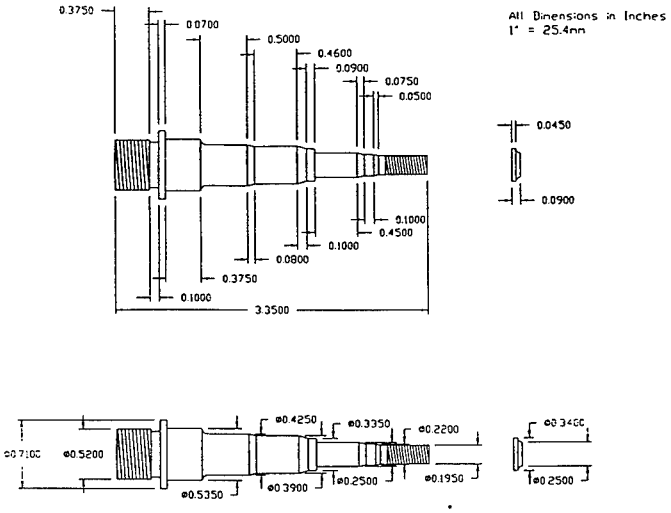


Figure 1a. Drawing of the Bicycle Pedal Spindle

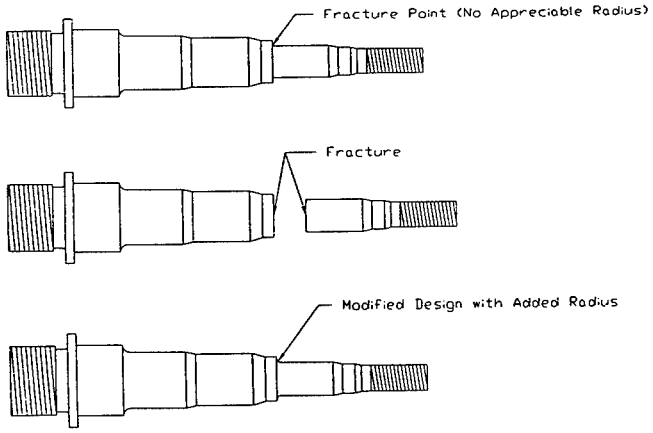


Figure 1b. Fracture Point of the Bicycle Pedal Spindle

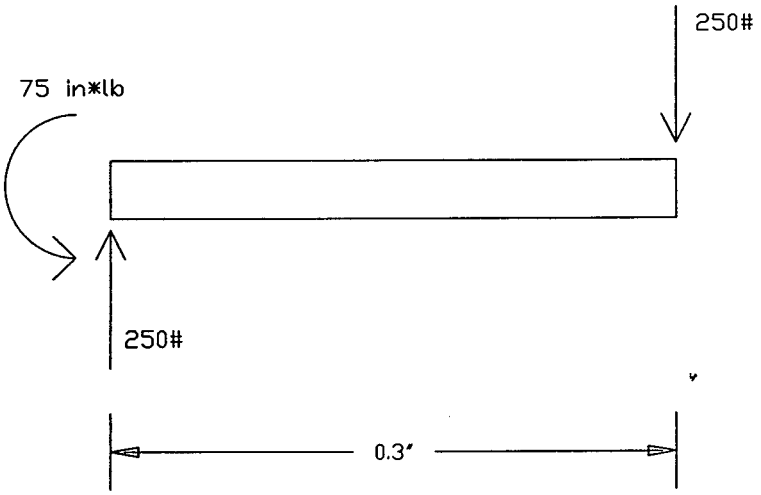


Figure 2. Free Body Diagram of Bicycle Pedal Spindle

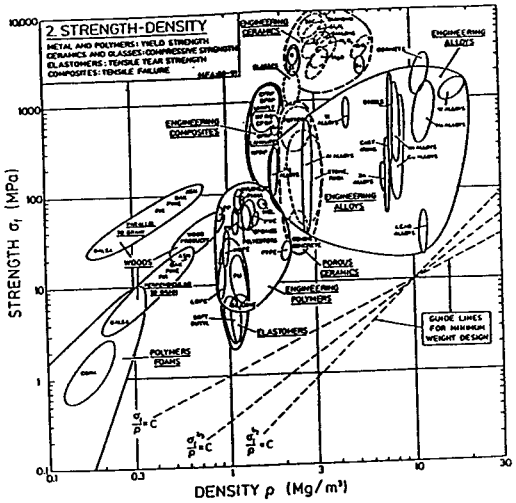


Figure 3. Strength vs. density for metals and composites. Strength is the yield stress for metals and the tensile strength for composites [4].

A GRAPHICAL APPROACH TO SPRING DESIGN

Martin L. Smith, Franklin E. Fisher and Omar S. Es-Said
Loyola Marymount University
Mechanical Engineering Department
Los Angeles, CA 90045

Abstract: Though spring stress strain relationships are well understood, selecting an optimum spring configuration for a given space envelope is not always easy. It is greatly simplified once a designer can readily visualize the trade-offs between material properties and physical dimensions, for a given manner of loading. A graphical approach, described here for metal close-wound torsional helical springs, aids such visualization.

Key Words: Spring; torsional; graphical; design; optimization

INTRODUCTION: Since a spring is often a lesser component in an overall design, the designer often wishes to constrain its size and end wire locations early in the design process. A graphical approach can help to quickly point to an optimum configuration knowing little more than a coil diameter, a spring length and a load or deflection. To effectively utilize this approach, a baseline design is selected from which alternate designs may be compared. It is also assumed that the needed spring supplies a minimum installed moment, M_0. The maximum operating moment will of course vary considerably from this value.

Beyond the spring's configuration, the designer must judge the price/performance trade off being made when selecting the spring wire. Herein, material performance will be assessed based upon the maximum storable elastic energy per unit wire volume.

In the method proposed here, these two items, material selection and physical configuration are performed in two separate steps.

NOMENCLATURE:

Variables:

- a = wire cross section axial length (parallel to spring axis)
- d = diameter
- k = stiffness
- t = wire cross section transverse length (perpendicular to spring axis)
- A = constant, characteristic of wire shape
- C = D_{av}/d or D_{av}/t
- D = coil diameter
- E = Young's modulus
- $\overline{E}$ = stored energy per unit volume of spring material uncorrected for curvature effects
- I = second area moment
- K_B = bending stress factor to account for spring winding curvature

L = coil length
LF = load factor (= $1/SF$)
M = moment required to deflect the spring
N = total number of coils, unadjusted
N' = number of coils, adjusted to account for constrained end wires
S = failure stress
SF = factor of safety
T = angular deflection
Δ = total angular deflection during operation, $T_{max} - T_{min}$
σ = tensile or compressive stress

Subscripts:
0 — baseline design
alt — alternating
av — average
e — fatigue endurance
f — failure
id — inside diameter
mean — mean
max — maximum
min — minimum
od — outside diameter
op — operating
rd — round cross section
sq — square or rectangular cross section
ult — ultimate
y — yield

MATERIAL SELECTION: Neglecting coil curvature effects, a wire's stored elastic energy per unit volume at failure (circular wire cross section) is [1]:

$$\overline{E} = \frac{\sigma_f^{\,2}}{8E} \text{ (round wire)} \qquad (1)$$

$\overline{E}$ for rectangular wire is one third higher. Table I shows $\overline{E}$ for common iron, nickel and copper based spring wire alloys. By balancing $\overline{E}$ against material cost, the designer can select the best material for a considered application.

Table I
Specific Energy Storage Capacity for Round Spring Wire

Material	E [2] Gpa (10^6 psi)	σ ‡ Mpa (ksi)	$\overline{E}$ J/cm^3 (ft-lbf/in^3)
Carbon Steels:			
Music Wire	207 (30)	1662 (241)	1.67 (20.2)
Hard Drawn	207 (30)	1324 (192)	1.06 (12.8)
Oil Tempered	207 (30)	1344 (195)	1.09 (13.2)
Stainless Steels:			
302	193 (28)	1276 (185)**	1.05 (12.7)
High Alloy Steels			
Chrome Silicon	207 (30)	1793 (260)	1.94 (23.5)
Chrome Vanadium	207 (30)	1517 (220)	1.39 (16.8)
Copper-Based Alloys:			
Phosphor Bronze	103 (15)	896 (130)	0.97 (11.7)
Beryllium Copper	128 (18.5)	1172 (170)	1.35 (16.3)
Spring Brass	110 (16)	120 (827)	9.4 (0.78)
Nickel-Based Alloys			
Inconel¥	214 (31)	1172 (170)	0.80 (9.7)
Inconel-X¥	214 (31)	1310 (190)	1.00 (12.1)
Inconel-X£	214 (31)	1517 (220)	1.35 (16.3)

‡ for ∅4.88mm (∅0.192in) wire — except as noted [3]
** for ∅5.26mm (∅0.207in) wire — except as noted [4]
¥ Spring temper
£ After aging

SPRING PERFORMANCE RELATIONSHIPS: A torsionally loaded helical spring's stiffness is [5]:

$$k_{rd}(d,N)=\frac{Ed^4}{10.8ND_{av}}$$

$$k_{sq}(a,t,N)=\frac{Eat^3}{6.6ND_{av}}$$

These two relationships can be simplified as follows:

$$k(I,N)=\frac{AEI}{2ND_{av}} \qquad (2)$$

where A is 3.77 and 3.63 respectively for wire of round and rectangular cross sections.

The maximum stress is at the ID [6][7]:

$$\sigma_{max} = \frac{M_{max}(t/2)}{I} K_B \tag{3}$$

or, expressed alternatively

$$\sigma_{max} = \sigma_{min} + \sigma_{op} = \frac{\left[M_0 + k\Delta\right](t/2)}{I} K_B \tag{4}$$

where [8]:

$$K_{Brd} = \frac{4C-1}{4C-4}$$

$$K_{Bsq} = \frac{4C}{4C-3}$$

and:

$$C \equiv D_{av} / d \text{ or } D_{av} / t$$

The deflection, T is simply equal to the spring moment divided by the stiffness:

$$T = \frac{M}{k}$$

Often, the end wire angular location may be constrained. Given that N equals the number of coils of an *undeflected* spring, to maintain a certain moment M_0 at a constrained angular position, a small adjustment must be made to account for differences in preload deflection when comparing designs For a baseline design having N_0 turns and minimum deflection T_0, the corrected N value is given by:

$$N' = N_0 - T_{min} + T_{0min} \tag{6}$$

Expressions for k, K_B, T and N' and σ are summarized in Table II and Table III for round and rectangular wire.

CYCLIC LOADING CONSIDERATIONS: The modified Goodman relation [9] is useful to assess failure limits for springs subject to cyclic loading:

$$\frac{1}{SF} = \frac{\sigma_{mean}}{S_{ult}} + \frac{\sigma_{alt}}{S_e}$$

Table II **Spring performance data expressed in terms of D_{av}, D_{id}, and D_{od}**
(wire of circular cross section)

	D_{id}	D_{av}	D_{od}
$k(d,N)$	$\frac{Ed^4}{10.8N\left(D_{id}+d\right)}$	$\frac{Ed^4}{10.8ND_{av}}$	$\frac{Ed^4}{10.8N\left(D_{od}-d\right)}$
$K_B(d)$	$\frac{4D_{id}+3d}{4D_{id}}$	$\frac{4D_{av}-d}{4\left(D_{av}-d\right)}$	$\frac{4D_{od}-5d}{4\left(D_{od}-2d\right)}$
$T_{min}(d,L)$	$\frac{128M_0L\left(D_{id}+d\right)}{3.77E\pi d^5}$	$\frac{128M_0LD_{av}}{3.77E\pi d^5}$	$\frac{128M_0L\left(D_{od}-d\right)}{3.77E\pi d^5}$
$N(d,N)$	$\left[1-\frac{128M_0\left(D_{id}+d\right)}{3.77E\pi d^4}\right]N+T_{0min}$	$\left[1-\frac{128M_0\left(D_{av}\right)}{3.77E\pi d^4}\right]N+T_{0min}$	$\left[1-\frac{128M_0\left(D_{od}-d\right)}{3.77E\pi d^4}\right]N+T_{0min}$
$\sigma_{min}(d)$	$\frac{32M_0}{\pi d^3}K_B$		
$\sigma_{op}(d,L)$	$\frac{3.77Ed^2\Delta}{4L\left(D_{id}+d\right)}K_B$	$\frac{3.77Ed^2\Delta}{4LD_{av}}K_B$	$\frac{3.77Ed^2\Delta}{4L\left(D_{od}-d\right)}K_B$
$\sigma_{max}(d,L)$	$\sigma_{min}(d)+\sigma_{op}(d,L)$		
$LF_{max}(d,L)$	$\frac{1}{2}\left[\frac{2\sigma_{min}(d)+\sigma_{op}(d,L)}{S_{ult}}+\frac{\sigma_{op}(d,L)}{S_e}\right]$		

Table III **Spring performance data expressed in terms of D_{av}, D_{id}, and D_{od}**
(wire of rectangular cross section)

	D_{id}	D_{av}	D_{od}
$k(a,t,N)$	$\frac{Eat^3}{6.6N\left(D_{id}+t\right)}$	$\frac{Eat^3}{6.6ND_{av}}$	$\frac{Eat^3}{6.6N\left(D_{od}-t\right)}$
$K_B(t)$	$\frac{4\left(D_{id}+t\right)}{4D_{id}+t}$	$\frac{4D_{av}}{4D_{av}-3t}$	$\frac{4\left(D_{od}-t\right)}{4D_{od}-7t}$
$T_{min}(a,t,L)$	$\frac{24M_0L\left(D_{id}+t\right)}{3.63Ea^2t^3}$	$\frac{24M_0L\left(D_{av}\right)}{3.63Ea^2t^3}$	$\frac{24M_0L\left(D_{od}+t\right)}{3.63Ea^2t^3}$
$N'(a,t,N)$	$\left[1-\frac{24M_0\left(D_{id}+t\right)}{3.63Eat^3}\right]N+T_{0min}$	$\left[1-\frac{24M_0\left(D_{av}\right)}{3.63Eat^3}\right]N+T_{0min}$	$\left[1-\frac{24M_0\left(D_{od}-t\right)}{3.63Eat^3}\right]N+T_{0min}$
$\sigma_{min}(a,t)$	$\frac{6M_0}{at^2}K_B$		
$\sigma_{op}(a,t,L)$	$\frac{3.63Eat\Delta}{4L\left(D_{id}+t\right)}K_B$	$\frac{3.63Eat\Delta}{4LD_{av}}K_B$	$\frac{3.63Eat\Delta}{4L\left(D_{av}-t\right)}K_B$
$\sigma_{max}(a,t,L)$	$\sigma_{min}(a,t)+\sigma_{op}(a,t,L)$		
$LF_{max}(a,t,L)$	$\frac{1}{2}\left[\frac{2\sigma_{min}(a,t)+\sigma_{op}(a,t,L)}{S_{ult}}+\frac{\sigma_{min}(a,t)}{S_e}\right]$		

or expressed alternatively:

$$LF = \frac{1}{SF} = \frac{1}{2}\left[\frac{2\sigma_{min} + \sigma_{op}}{S_{ult}} + \frac{\sigma_{\min}}{S_e}\right] \tag{7}$$

Eq. (7) can be used in place of Eqs. (3) or (4) when considering cyclically loaded springs.

S_e varies depending upon the material, but for most steels, it is roughly related as follows to tensile strength [10]:

$$S_e = \begin{cases} 0.504\sigma_{ult} & \sigma_{ult} \leq 200 \text{ ksi (1400 MPa)} \\ 100 \text{ ksi (700 MPa)} & \sigma_{ult} > 200 \text{ ksi (1400 MPa)} \end{cases}$$

GENERATING THE PLOTS: Objective functions σ_{max} and LF are themselves functions of the wire cross section (a and t or d) and spring length, L. Given a baseline spring length L_0 with N_0 turns, we can generate two families of curves on a plot of wire size (a or d) vs. the relevant objective function. The first family of curves has the argument L in the objective function set equal to L_0 plus or minus one or two turns of wire (e.g. $L = L_0 - 2a$, $L_0 - a$, L_0, $L_0 + a$ and $L_0 + 2a$). The second family of curves has L set equal to the quantity N0′ plus or minus several turns multiplied by the wire size: (e.g. $L = (N_0 - 2)a$, $(N_0 - 1)a$, $N_0\, a$, $(N_0 + 1)a$, and $(N_0 + 2)a$). The baseline stress or load factor and wire size can be found at the point where the $L = L_0$ and $L = N_0\, a$ plots intersect (Fig. 1). Thus plotted, the designer can quickly find the wire size and coil count which minimize σ_{max} or LF.

SUMMARY: By balancing the storable elastic energy per unit wire volume (material performance) for several spring materials against a material's value, a suitable spring material can be easily selected.

Once selected, the plots described above can be generated and modified quickly and conveniently with the help of a computer graphing program. Once plotted, it takes only a quick look at the plot and a table showing standard spring wire sizes to then specify the best spring to accomplish the needed task. This same approach could be similarly applied to other spring types as well.

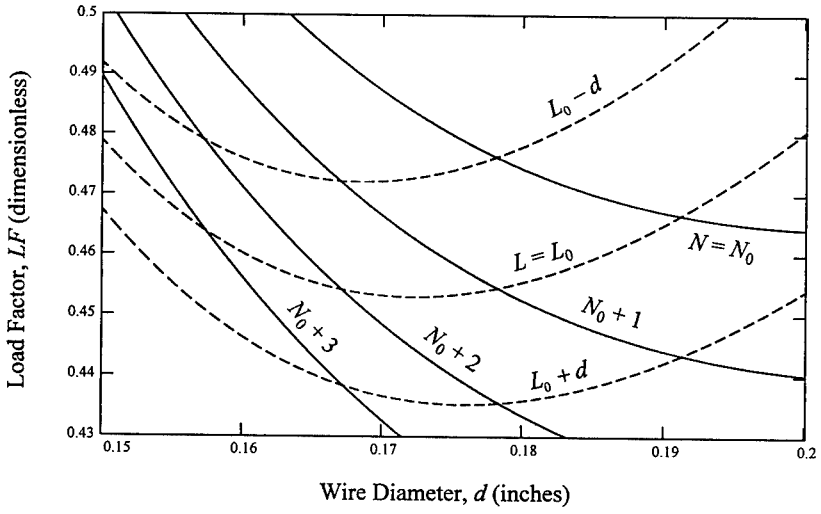


Figure 1 — Sample Performance Optimization Plot

REFERENCES

[1] Associated Spring, Design Handbook: Engineering Guide To Spring Design, 1987 ed., Barnes Group, (1987), p. 10.

[2] *Ibid.*, p. 12.

[3] Spring Manufacturers Institute (SMI), Handbook of Spring Design, SMI, Rolling Meadows, IL (1993), pp. 44-46.

[4] *Ibid.*

[5] *Ibid.*, p. 24

[6] Timoshenko, S., Strength of Materials, Vol. 1, 2nd ed., D. Van Nostrand, New York, (1940), p. 92.

[7] Associated Spring, *Op. Cit.*, pp. 62-63.

[8] *Ibid.*

[9] J. E. Shigley and C. R. Mischke. Mechanical Engineering Design, 5th ed., McGraw-Hill, New York, (1989), p. 299.

[10] *Ibid.*, p. 278.

INFANT CRIB FAILURE ANALYSIS CASE STUDY

Dennis B. Brickman
Triodyne Inc.
5950 W. Touhy Ave.
Niles, IL 60714

Ralph L. Barnett
Illinois Institute of Technology
Dept. of Mechanical and Aerospace Engineering
Chicago, IL 60616

Abstract: The evaluation of infant cribs by the ASTM test protocol is enhanced by random vibration testing. This testing adds a quantitative component to the conventional go/no-go criteria and produces a hierarchy of weakest links. Furthermore, this failure analysis technique is shown to discriminate among subtle differences in the structural design of the cribs.

Key Words: Crib; failure analysis; testing

ACCIDENT DESCRIPTION: An infant crib was alleged to cause a child's death when the lower fasteners on the stationary side rail illustrated in Fig. 1a and Fig. 1b became undone allowing an escape corridor to develop. The stationary side rail was against the wall and the child wedged his head between the mattress and the side rail causing his strangulation. It was proposed that the grommet nuts and screws were inadequately designed because they lacked a locking device and redundancy. The idea was the racking environment of the crib would loosen the unrestrained fasteners and allow them to dislodge. It was alleged by the child's father that one or both of the stationary side rail lower fasteners were missing and when the crib was subsequently located away from the wall, the side rail could swing open about the top fasteners in an unimpeded fashion. Furthermore, the child's father saw a hex nut and screw laying directly under the crib after the accident. However, the lower fasteners were not produced with the accident crib.

ACCIDENT STATISTICS: According to a U.S. Consumer Product Safety Commission (CPSC) Special Report on Structural and Hardware Failure of Cribs [1], 167 incidents of failure of crib hardware or other structural components of cribs were reported to the CPSC for the period from January 1, 1980 through December 31, 1983. After a detailed analysis, the CPSC categorized these 167 incidents by hazard pattern type as shown in Table 1. Additional incidents of a similar nature have been reported by the CPSC subsequent to this period. In response to the accident data collected by the CPSC, a new standard, ASTM F 1169-88 [2], was developed by the Juvenile Products Manufacturing Association (JPMA) and approved by a vote of the American Society for Testing and Materials (ASTM) members to establish test requirements to address failures associated with mattress support hardware, glued or bolted connections, dropside latches, and dislodgment of teething rails. This safety

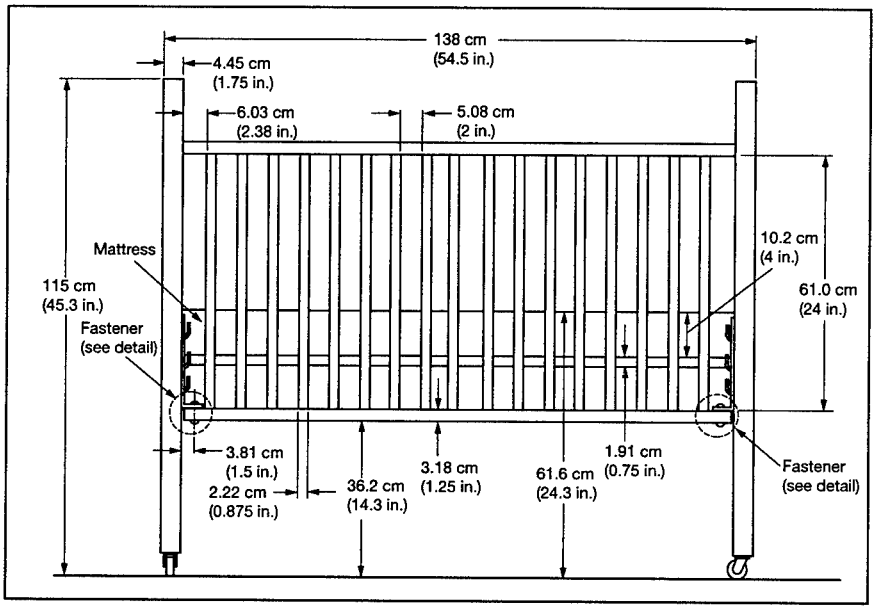


Fig. 1a Infant Crib

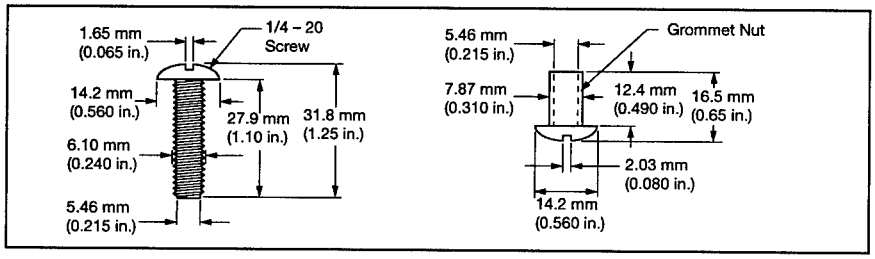


Fig. 1b Infant Crib Fastener Detail

Table 1 Incidents Related to Lack of Structural Integrity of Cribs, by Hazard Pattern Category and Severity of Accident (January 1, 1980 – December 31, 1983)

Hazard Pattern Category	Severity of Accidents					
	TOTAL	Deaths	Near-Miss Entrapments	Medically-Treated Injuries	Minor Injuries	No Injury
TOTAL	167	34	22	7	47	57
Mattress-Support Hardware Failure	32	2	11	–	11	8
Attachment Hardware Failure	18	5	4	–	2	7
Structural/Material Failure	47	3	3	2	14	25
Dropside Mechanism Failure	23	–	–	3	11	9
Maintenance, Repair/Assembly	15	13	–	–	2	–
Portable Crib Failure	17	1	2	2	5	7
Hardware Problem Unclear	15	10	2	–	2	1

Source: U.S. Consumer Product Safety Commission, Directorate for Epidemeology.

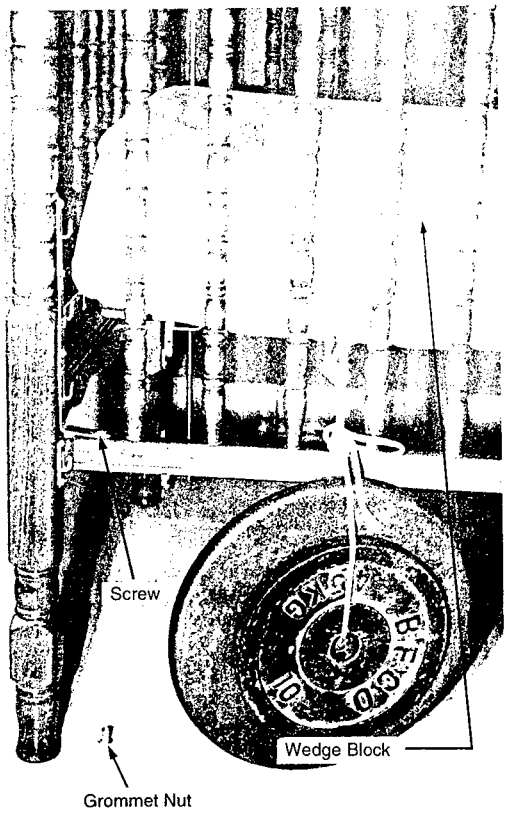


Fig. 2 ASTM Vertical Wedge Block Test Using Screw Only

standard also addresses incidents associated with poor maintenance or assembly by specifying requirements for the contents of instructional literature that must accompany cribs.

ASTM TESTING:

Vertical Wedge Block Testing: Vertical wedge block testing originally developed for bunk beds was performed on an exemplar crib according to section 5 of ASTM F 1427-92 Standard Consumer Safety Specification for Bunk Beds [3]. The specified wedge block was inserted tapered side downward into the gap between the interior crib stationary side rail structure and the edge of the mattress. A 200 N (45 lb) force was gradually applied downward and was sustained for a period of one minute. This test protocol was conducted once utilizing the original grommet nut and screw combination, once with only the screw inserted in the top of the left bottom stationary side rail, and once with the grommet nut only inserted in the top of the left bottom stationary side rail.

Performance requirements for the vertical wedge block testing are given by section 4 of ASTM F 1427-92 as follows [3]: "There shall be no gaps between the interior bed structure and the edges of the mattress and foundation that will permit complete passage of the wedge block." The wedge block did not completely pass between the crib stationary side rail structure and the mattress during any of the three trials. Figure 2 depicts the vertical wedge block test condition using only the screw with the grommet nut on the floor. It should be noted that the original grommet nut and screw design is redundant with respect to the orientation of the fastener. That is, using only a grommet nut on the top or only a screw on the top will lead to compliance with the ASTM vertical wedge block test.

Vertical Impact Testing: Vertical impact testing was conducted on an exemplar crib according to the test methods specified in section 5 of ASTM F 1169-88 Standard Specification for Full Size Baby Crib [2]. A 343 mm (13.5 in.) diameter 20.4 kg (45 lb) weight was allowed to free fall 152 mm (6 in.) on to the upper surface of a foam pad at a rate

of 4 seconds per cycle as shown in Fig. 3. Five hundred cycles were conducted at the geometric center of the mattress area and 100 cycles were conducted at each of two diagonally opposite corners centered 230 mm (9 in.) from the crib sides forming the corner. The test protocol was carried out using the original grommet nut and screw combination and also using a hex nut and 1/4-20, 50.8 mm (2 in.) long replacement screw inserted in the left bottom stationary side rail.

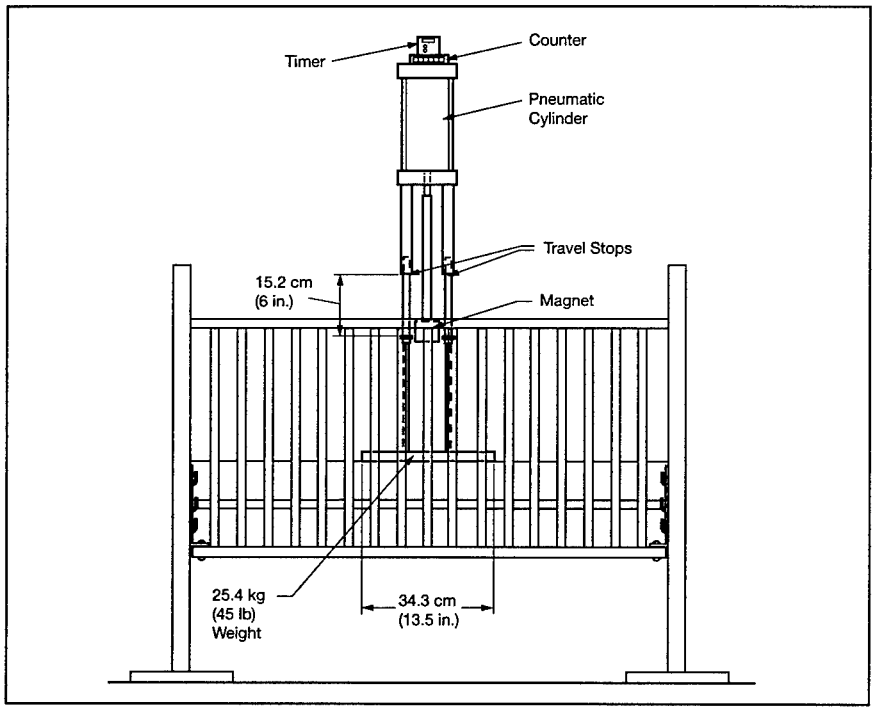


Fig. 3 ASTM Vertical Impact Test Set Up

Performance requirements for the vertical impact testing are given by section 5.5 of ASTM F 1169-88 as follows [2]: "Components attached by screws shall not have separated by more than 0.04 in. (1 mm) upon completion of testing." The components attached by screws did not separate at all upon completion of the testing using the original or replacement hardware. Figure 4 depicts the condition of the original grommet nut and screw combination upon completion of the vertical impact testing.

Stationary Side Cyclic and Static Testing: Stationary side cyclic and static testing was performed on an exemplar crib according to the protocol specified in section 6 of ASTM F 1169-88 [2]. For the cyclic test, an 11.3 kg (25 lb) weight was allowed to free fall 76 mm (3 in.) 50 times at a rate of 4 seconds per cycle such that it impacted upon a 9 mm (0.375 in.) rubber pad located on the top surface of the bottom rail between two adjacent slats as near the center of the rail as possible as shown in Fig. 5. Upon completion of the cyclic test,

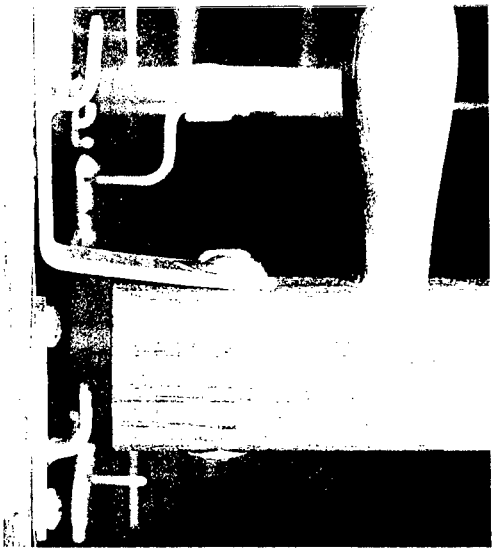


Fig. 4 Original Grommet Nut and Screw after ASTM Vertical Impact Test

a static load of 45.4 kg (100 lb) was applied at the point of impact testing gradually within a period of 5 seconds and maintained for an additional 30 seconds. The testing was first conducted using the original grommet nut and screw combination and then using a hex nut and 1/4-20, 50.8 mm (2 in.) long replacement screw inserted in the left bottom stationary side rail.

Crib side testing requirements are given by section 6.9 of ASTM F 1169-88 as follows [2]: "Components attached by screws shall not have separated by more than 0.04 in. (1 mm) upon completion of testing." The components attached by screws did not separate at all upon completion

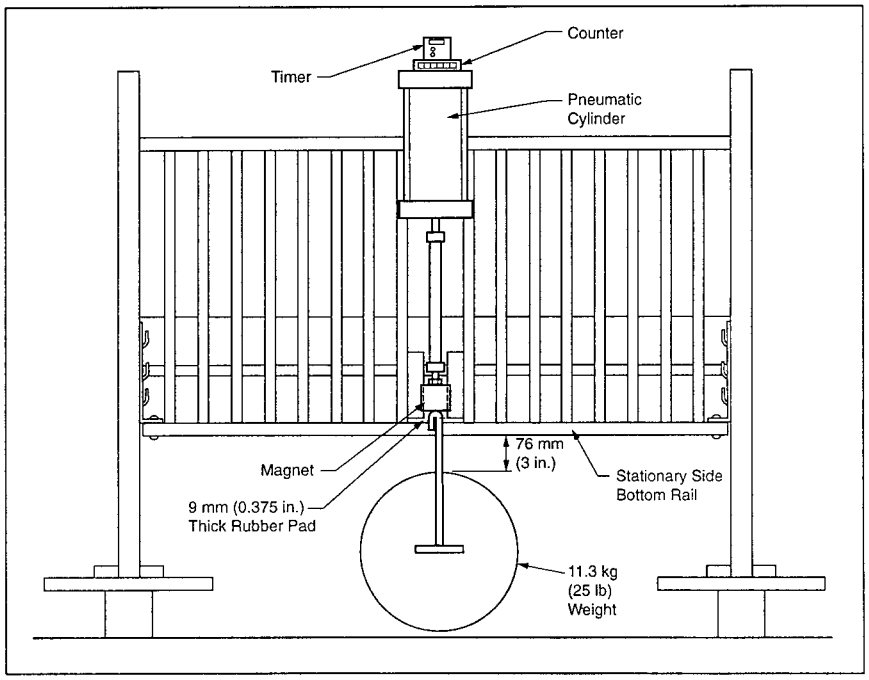


Fig. 5 ASTM Stationary Side Cyclic Test Setup

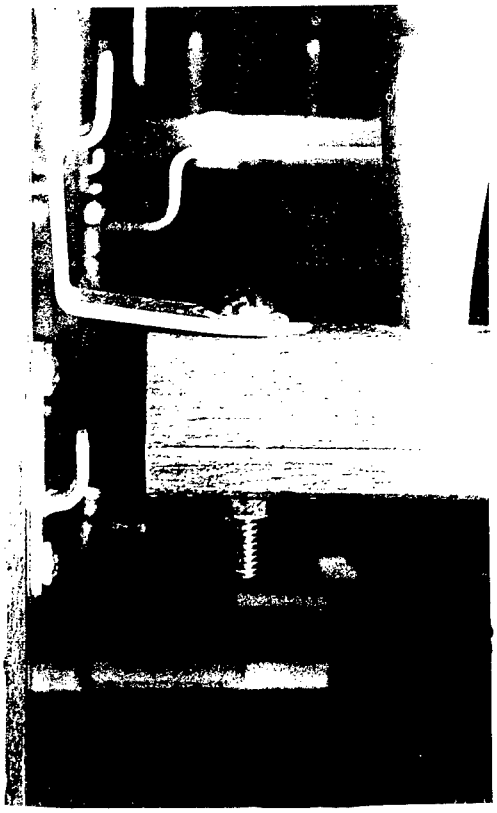


Fig. 6 – Hex Nut and Replacement Screw after Stationary Side Cyclic and Static Testing

of the testing using the original or replacement hardware. Figure 6 shows the tightened status of the hex nut and replacement screw upon completion of the stationary side cyclic and static testing.

SHAKE TABLE TESTING: Since the exemplar crib met the requirements of the ASTM test protocol, this suggests that the crib design was not defective. To challenge the notion that compliance with the ASTM standard establishes the product's safety, it was suggested that the ASTM loading environment was not sufficiently intense to simulate the reasonably foreseeable loading history. To rebut this attack, the failure analysis continued applying the strongest link principle approach with shake table testing [4]. According to the strongest link principle, if a challenged design feature on an exemplar is unaffected by a simulated or accelerated loading program which causes the exemplar to undergo changes which differ from the artifact, then failure of the artifact is not proximately caused by the challenged design feature.

Description of Equipment: Figure 7 illustrates the shake table which consists of a platen which is vertically driven at three locations by independently controlled pneumatic cylinders with a maximum stroke of 20.3 cm (8 in.). The stroke time history of each cylinder is computer controlled leading to motions that involve vertical translation and rotations about two axes in a horizontal plane. The shake table does not produce rotation about a vertical axis or horizontal translations. The specifications for the shake table and its control system are described by Brickman and Barnett in the Strongest Link Principle [4].

Testing Protocol: An exemplar crib was mounted on the shake table at the four legs with an 11.3 kg (25 lb) weight centered on the mattress to reflect the mass of the deceased infant. A severe shaking environment was developed which led to the failure and virtual destruction of the crib. This test protocol was conducted once using the original grommet nut and screw combination, once using a hex nut and 1/4-20, 50.8 mm (2 in.) long replacement screw inserted in the left bottom stationary side rail, and once using a competitor crib design.

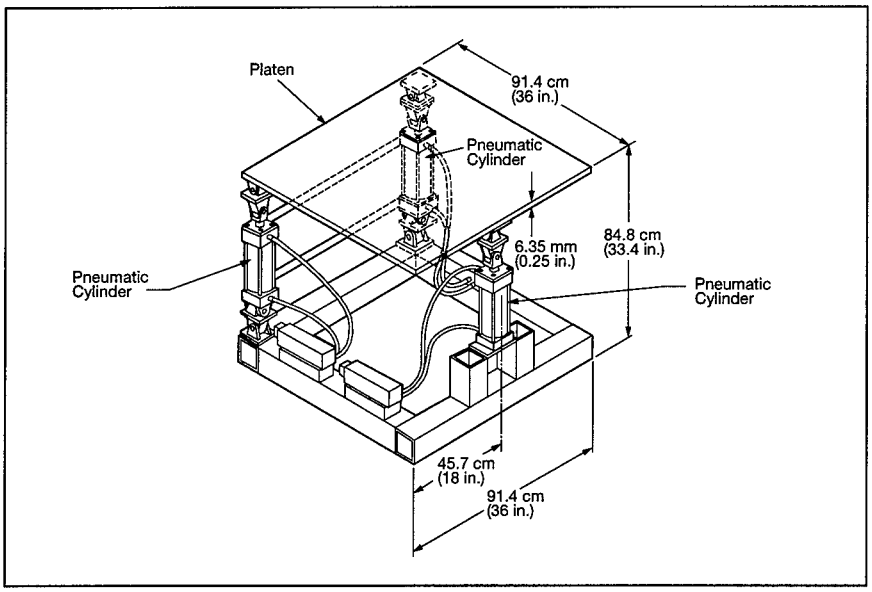


Fig. 7 Shake Table

Test Results:

Original Hardware: The following observations in sequencial order were made during the shake table testing with the original hardware:

1. The mattress support spring separated from the mattress spring hooks and fell to the floor.

2. The drop side rail lowered.

3. The short machine screw released the top left of the stationary side rail from the foot panel.

4. The short machine screw released the right drop side rod from the foot panel.

5. The original grommet nut and long machine screw fasteners on the bottom of the stationary side rail remained connected.

The conditon of the crib with the original hardware at the conclusion of the shake table testing is depicted in Fig. 8. The numerical call outs shown in Fig. 8, Fig. 9, and Fig. 10 correspond to the itemized test result descriptions given for the original hardware, replacement hardware, and competitor crib cases respectively.

Replacement Hardware: Observations made in sequential order during the shake table testing with the replacement hardware are as follows:

1. The mattress support spring separated from the mattress spring hooks and fell to the floor.

2. The drop side rail lowered.

3.The short machine screw released the right drop side rod from the foot panel.

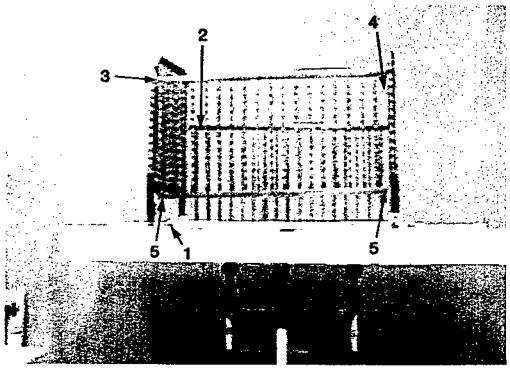


Fig. 8 Exemplar Crib with Original Hardware after Shake Table Testing

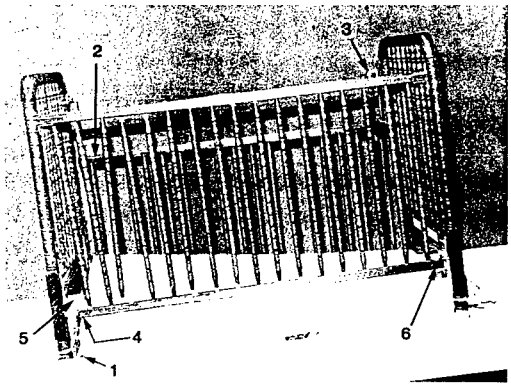


Fig. 9 Exemplar Crib with Replacement Hardware after Shake Table Testing

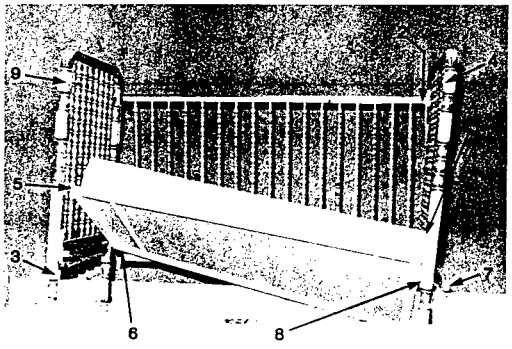


Fig.10 Competitor Crib after Shake Table Testing

4. The hex nut released from the replacement screw at the left bottom of the stationary side rail.

5. The replacement screw came out of the left bottom stationary side rail and remained in the left top bracket.

6. The original grommet nut and long machine screw fasteners on the right bottom stationary side rail remained connected.

The condition of the crib with the replacement hardware at the conclusion of the shake table testing is shown in Fig. 9.

Competitor Crib: The following observations in sequential order were made during the shake table testing of the competitor crib:

1. The right rear mattress support separated from the headboard.

2. The right top drop rail fastener failed.

3. The bottom left stationary side rail separated from the footboard.

4. The top right stationary side rail separated from the headboard.

5. The left front mattress support separated from the footboard.

6. The bottom left drop rail separated from the footboard.

7. The headboard failed and separated.

8. The bottom right stationary side rail separated from the headboard.

9. The top left stationary side rail separated from the footboard and the stationary side rail released from the crib.

The condition of the competitor crib at the conclusion of the shake table is depicted in Fig. 10.

WARNINGS AND INSTRUCTIONS: At the time of manufacture of the subject crib, CPSC Part 1508 - Requirements for Full-Size Baby Cribs [5] contained specifications for assembly instructions and warnings associated with cribs. In accordance with the CPSC requirements, the subject crib contained the following on-product warnings and instructions:

1. Caution: Any mattress used in this crib must be at least 27-1/4 inches by 51-5/8 inches with a thickness not exceeding 6 inches.

2. Children should not be allowed to climb on crib.

In addition, the written instructions that accompanied the subject crib contained the following applicable warnings and instructions:

1. For the safety of your child and your own piece of mind, be sure you follow the assembly instructions exactly.

2. Lay out and check the parts you received with figure 1 and the following parts list.

3. Attach the stationary side rail bottom to the top brackets above it with long screws and nuts.

4. Check regularly to see that all screws are tight.

5. If you need parts write to customer service.

It should be noted that although the child's father never received the written instructions, he did not claim to have had any difficulty in assembling the subject crib. On the other hand, the child's father did not follow the warnings and instructions on the subject crib, allowing his other child to climb and jump on the crib and using a mattress which was smaller than the specified dimensions.

In January of 1989, a new standard was published and designated as ASTM F 1169-88, Standard Specification for Full Size Baby Crib. Sections 10 and 11 of this standard contain additional labeling and instructional literature requirements to address incidents associated with poor maintenance or assembly reported in the field for all full size baby cribs [2]. In response to this new standard, the manufacturer of the subject crib updated its written instructions to include the warnings and instructions specified by ASTM. The pertinent additions are:

1. Before each usage or assembly, inspect crib for damaged hardware, loose joints, missing parts or sharp edges.

2. Do not use crib if any parts are missing or broken. Ask your dealer or write for replacement parts and instructional literature if needed. Do not substitute parts.

CONCLUSIONS: The ASTM test requirements for the crib stationary side rail describe a go/no-go program [6] which gives information on the total behavior of the system, but provides no information about the weak links. In addition, the ASTM go/no-go criteria provide no quantitative information regarding the level by which a particular crib design exceeds the specified performance requirements. The accelerated loading program, executed through shake table testing, facilitates an ordering of weak links in the system. According to the strongest link principle, the challenged design feature cannot be a proximate cause of the subject crib failure if it is preceded by weak link failures not present on the artifact.

Furthermore, the random vibration method produced by the shake table testing reveals subtle differences among fasteners and distinguishes between proper and improper installation states such as the vertical orientation of a screw head. Because of available symmetry, comparisons can be made between hardware on the left and right sides of the stationary side rail, making it very clear which is better in a random vibration load environment.

The results of the shake table testing on an exemplar crib indicate the virtual destruction of the crib with no concomitant loosening of the original stationary side rail lower fasteners. Loading which is sufficient to compromise the lower fasteners will result in the signature fractures experienced in the shake table testing program, but were absent in the artifact. This establishes that the original grommet nut and screw design was not a proximate cause of the subject accident. Alternative explanations may be failure to install the lower fasteners, improper tightening of the fasteners, utilization of loose replacement hardware, removal or loosening of the fasteners by unknown parties and the like. Shake table testing of the competitor crib revealed failure modes that were thought to be unachievable due to redundancy. Through the shake table testing, it was discovered that a single failure of the competitor crib can create an unsafe condition.

The CPSC and ASTM standards have codified warnings and instructions for full size baby cribs. According to the principle of uniform safety [7], the resulting technology transfer will serve the general community of users who will be presented with the same information no matter which product is looked at. However, these warnings and instructions address general problems that may not necessarily be present on a given design. Therefore, a given crib may be mandated to convey warnings and instructions which imply a problem that does not exist.

Considering the subject crib design, the original stationary side rail grommet nut and screw combination is the strongest link in the system and is capable of surviving a loading environment which completely destroyed a competitor crib.

REFERENCES:

1. Rutherford, G.W., Special Report: Structural and Hardware Failure of Cribs, U.S. Consumer Product Safety Commission, Division of Hazard Analysis (EPHA), Directorate for Epidemiology, Washington, D.C., Mar. 1984.
2. "Standard Specification for Full Size Baby Crib," ASTM F 1169 - 88, American Society for Testing and Materials, Philadelphia, PA, approved Aug. 26, 1988.
3. "Standard Consumer Safety Specification for Bunk Beds," ASTM F 1427 - 92, American Society for Testing and Materials, Philadelphia, PA, approved Sep. 22, 1992.
4. Brickman, D.B. and Barnett, R.L., "Strongest Link Principle," ASME 94-WA/DE-2, American Society of Mechanical Engineers, 1994.
5. "Part 1508 - Requirements for Full-Size Baby Cribs," 16 CFR 1508, U.S. Consumer Product Safety Commission, Washington, D.C., Jan. 1, 1985, pp. 409-415.
6. Brickman, D.B. and Barnett, R.L., "Quantification Versus Go/No-Go Criteria," Reliability,Stress Analysis, and Failure Prevention - 1993, R.J. Schaller ed., American Society of Mechanical Engineers, DE-Vol. 55, 1993, pp. 9-15.
7. Barnett, R.L., "The Principle of Uniform Safety," Safety Engineering and Risk Analysis, D.W. Pyatt ed., American Society of Mechanical Engineers, SERA-Vol. 2, 1994, pp. 141-147.

On the Corrosion of AISI 304 and Galvanized Steel at a Meat Packaging Company

David E. French, Ignacio Hernandez, Richard W. Aston, Justin M. Busse
Martin L. Smith, Joseph Foyos and Omar S. Es-Said

Mechanical Engineering Department
Loyola Marymount University
Los Angeles, CA

Abstract: A meat packaging and processing company had a history of corrosion problems in their smoke house facility. The rate of corrosion significantly increased shortly after an exhaust ventilator was replaced with a smaller unit and the liquid smoke solution was changed. This study determined that the new liquid smoke solution was the cause of the increased corrosion rate. Several strategies for reducing the corrosion rate are recommended and evaluated.

Key Words: Corrosion; galvanized steel; AISI 304 stainless steel; liquid smoke solution; meat packaging company.

INTRODUCTION: A local meat packing and processing company has an elaborate layout for adding their famous "Hickory-smoke" flavor to their hot dogs. The process for adding this flavor is done in a facility called the smoke house. This facility had a history of corrosion problems. Recently, the corrosion rate increased significantly. This occurred shortly after the smoke house was put back in operation following a fire two years ago. After that fire, two changes were made, an exhaust ventilator was replaced (at a reduced size) and the liquid smoke solution was replaced. The corrosion problem was evident on the stainless steel panels and the galvanized steel chain. The stainless steel AISI 304 panels are 3.5" x 2.5" with a 1/8" thickness and the galvanized steel chain is 3,000' long and 1/4" thick. The corrosion is taking place in the smoke house where a liquid smoke solution is sprayed onto the sausages, after they are cooked and chilled, to provide a smoke flavor to the meat.

In zone 1 (Figure 1) stainless steel paneling is corroding rapidly and completely. This zone is at the end of the cooking area of the smoke house. In the same area, a stainless steel bolt was noted to be corroding at an accelerated rate. In zone 2, there is a fan that creates a "humidity barrier" between the heating area and the cooling area. This fan had never experienced a high rate of corrosion in the past, and has now corroded completely such that it had to be replaced. The third area of corrosion is the galvanized steel chain that holds the sausages and takes them through the smoke house. The steel chain is corroding at a rate which is twice the rate prior to the accident. Finally, in the Brine Chill Area (zone 4), heat exchanger coils are experiencing corrosion. This had previously not been a problem.

INVESTIGATION: The investigation included site visits and laboratory experiments. Table 1 describes the temperature variation and cooking cycle time within the smokehouse, (Figure 1).

Table 1. Temperature variation within the smoke house.

Location	Temperature
Pre-Heat Area	155°F for 15 minutes
Heating Area 1	175-190°F for 10 minutes
Heating Area 2	180-190°F for 15 minutes with steam (added 55% humidity)
Chill Area 1	40°F for 15 minutes
Brine Chill Area	25°F for 10 minutes
Chill Area 2	40°F for 1 minute

Though the reduction in size of the exhaust ventilation was significant (~50%), it did not explain the corrosion of the brine chiller coils nor the galvanized steel chain. It appeared that the new liquid smoke solution was the more probable cause, such that some new element(s) had been introduced that caused the accelerated corrosion rate.

Tests were performed on the galvanized steel chain. Pieces of the chain were placed in bottles, one bottle contained the old liquid smoke solution, another the new liquid smoke solution and a third with no solution (Figure 2). The third bottle contained no solution to test for the possibility of stress corrosion based solely on temperature change. The experiment was set up to simulate the temperature conditions at the smoke house. Also, pieces of galvanized steel chain were placed in two beakers containing the solutions, these were not heated nor cooled. This was done in an effort to examine the effect of direct chemical attack. The three bottles were heated for 12 hours at 185°F and then cooled at 28°F for another 12 hours. The experiment confirmed the initial hypothesis that some chemical(s) in the liquid smoke was responsible for causing the increased corrosion rate in the galvanized steel chain.

The experiment revealed that the galvanized steel pieces that were not exposed to a temperature change had no signs of corrosion. The piece that was heated and cooled with no liquid smoke added also showed no signs of corrosion. The galvanized steel piece in the old solution had very little signs of corrosion, but the galvanized steel piece that was in the new solution was completely corroded (Figures 3-6). This confirmed that the problem lies in the new liquid smoke solution. Next, the same experiment was done with the AISI 304 stainless steel. After one week of heating and cooling there were no signs of corrosion in the metal. Both liquid smokes were sent to a commercial laboratory* for analysis of pH and Chloride levels. The results are shown in Table 2 and they indicate no significant variations between both liquid smoke solutions.

* BC Analytical, 801 Western Ave., Glendale, CA 91201

Table 2. Laboratory Results.

Lab Results	pH	Chloride levels
Old Solution:	2.5	330 mg/L
New Solution:	2.6	290mg/L

To more nearly duplicate the corrosion environment, it was suggested that the samples be exposed to the liquid smoke solutions in a vapor form [1]. The AISI 304 and zinc galvanized steel samples were suspended in the glass containers above the smoke solutions, (Figure 7). After 72 hours of heating, both types of steel were examined. The results are shown in Table 3.

Table 3. Corrosion of Steel Samples.

Type of Steel	Visual Corrosion
Galvanized steel suspended above old liquid smoke solution:	very little corrosion evident
Galvanized steel suspended above new liquid smoke solution:	complete corrosion
AISI 304 suspended above old liquid smoke solution:	no corrosion evident
AISI 304 suspended above new liquid smoke solution:	complete corrosion

DESIGN SOLUTIONS: Based on the testing and analysis conducted, those methods for reducing the corrosion rate were suggested.

Design Concept A: REPLACE NEW LIQUID SMOKE WITH OLD LIQUID SMOKE. Definitely the simplest solution. Replace the new liquid smoke solution with the old solution. The change would be the least costly and have no change in the processing facility or operations.

Design Concept B: REPLACE AISI 304 AND ZINC GALVANIZED CHAIN WITH ALTERNATIVE METALS. Galvanized steel chain is problematic. Typically, zinc is used as a corrosion inhibitor in aqueous solutions in the pH range of 6.5 to 12. Highly acidic environments will corrode zinc rapidly. This is the problem in the smokehouse where the pH of either liquid smoke is around 2.5. The temperature ranges of the smoking process also create a corrosion problem for zinc. The rate of corrosion in zinc is highly dependent upon the temperature of the environment, [2]. The operating range of temperatures during the smoking process ranges from 28°F to 190°F (-16°C to 74°C). The temperature ranges from 155°F to 190°F (55°C to 74°C) for more than half of the entire smoking process. This range of upper temperatures is within the range of increased corrosion for zinc in water, [2]. Furthermore, acidic organic compounds are known to

readily corrode zinc. Due to the environment in the smokehouse the zinc galvanized steel chain should be eliminated and replaced with a AISI 316 stainless steel chain. With regards to the AISI 304 panel, this solution recommends its replacement with AISI 316 stainless steel. AISI 304 stainless steel is susceptible to corrosion attack by organic acids, [3]. The presence of acetic, citric, formic or lactic acid in the liquid smoke would be damaging to the AISI 304 stainless steel. The addition of molybdenum to the stainless steel, as is the case in AISI 316, has a significant effect in reducing the corrosion effects due to various acids, [4,5]. As outlined earlier., the AISI 304 heat exchanger is exhibiting corrosion in the cold worked coiled section, but not in the straight section of the tubes. This seems to be a classic stress corrosion cell. The material in the cold worked section of the heat exchanger is in a higher state of stress and therefore has a lower activation energy to the onset of corrosion. This, coupled with a more corrosive liquid smoke solution, appears to be the source of the corrosive attack. The corrosion rates increase with the increasing stress when AISI 304 is subjected to a chloride environment [6]. Increasing the content of Mo, Ni, and Cr in the alloy increases the life of the steel. Therefore, a material change to AISI 316 will improve corrosion resistance.

Design Concept C: ADDITION OF A PASSIVATION LAYER TO THE STAINLESS STEEL. Another reason for the increased corrosion with no increase in chloride ion concentration could be the absence of certain beneficial anions. The pitting potential of stainless steel in chloride solutions has been observed to be reduced by the presence of these beneficial anions [6], SO_4^{-2}, OH^-, CLO_3^-, CrO_4^{-2}, and NO_3^- all have corrosion inhibiting tendencies. The concentrations of various inhibiting anions could be tested for and may help slow the corrosion rate. These inhibitors work in many ways from stifling the anodic process to rendering the metal surface hydrophobic by adsorption of the inhibitor molecule. It is recommended by this solution that one of these inhibitors be utilized to create a passivation layer on top of the stainless steel.

EVALUATION AND RANKING: The evaluation of testing and possible solutions indicates that the company's decision to change their liquid smoke solution was a poor one. The new design concepts were ranked based upon the following criteria: cost effectiveness and effective corrosion resistance.

COST EFFECTIVENESS: The most cost effective solution presented was Design Concept A, all others required sizable financial investment. The company representatives could also subcontract a chemistry laboratory for a comprehensive analysis of the composition of both liquid smoke solutions and trace the undesirable element(s) of the new liquid smoke and, eventually, contact other suppliers to purchase a liquid smoke of a similar composition to the original liquid smoke solution.

EFFECTIVE CORROSION RESISTANCE: For this concept all solutions were very effective. Design Concept B seems to be the best. Design Concept C would require more testing to ensure it would work in the smokehouse environment and Concept A had a predictable outcome. With Concept A the corrosion would still occur, but at a much slower pace.

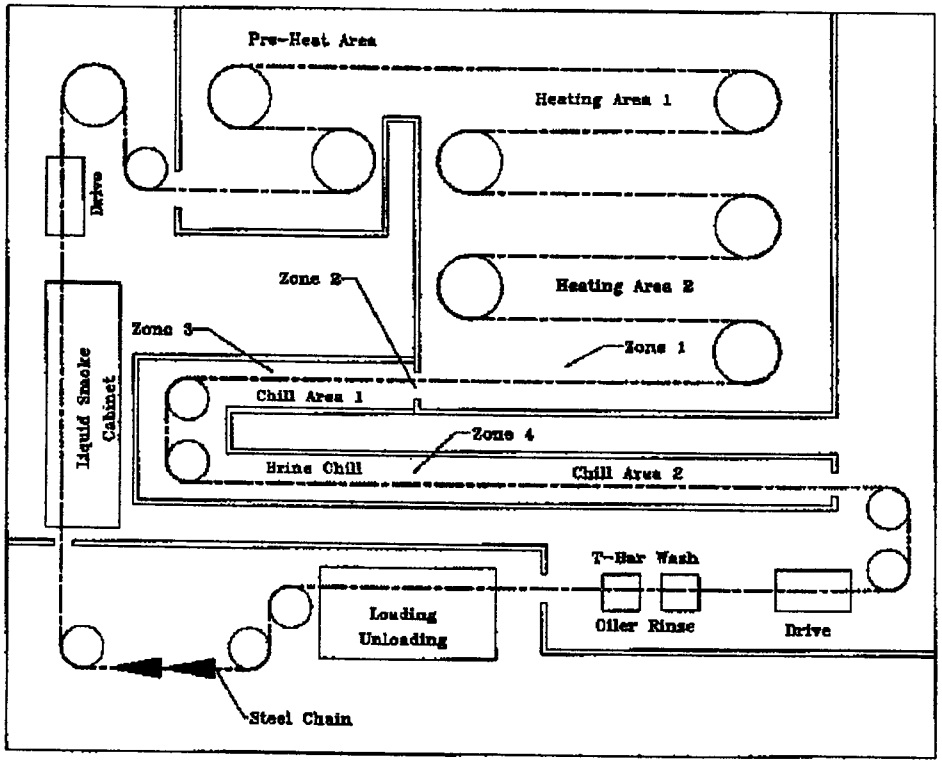


Figure 1: Schematic of the different zones of the Smoke House.

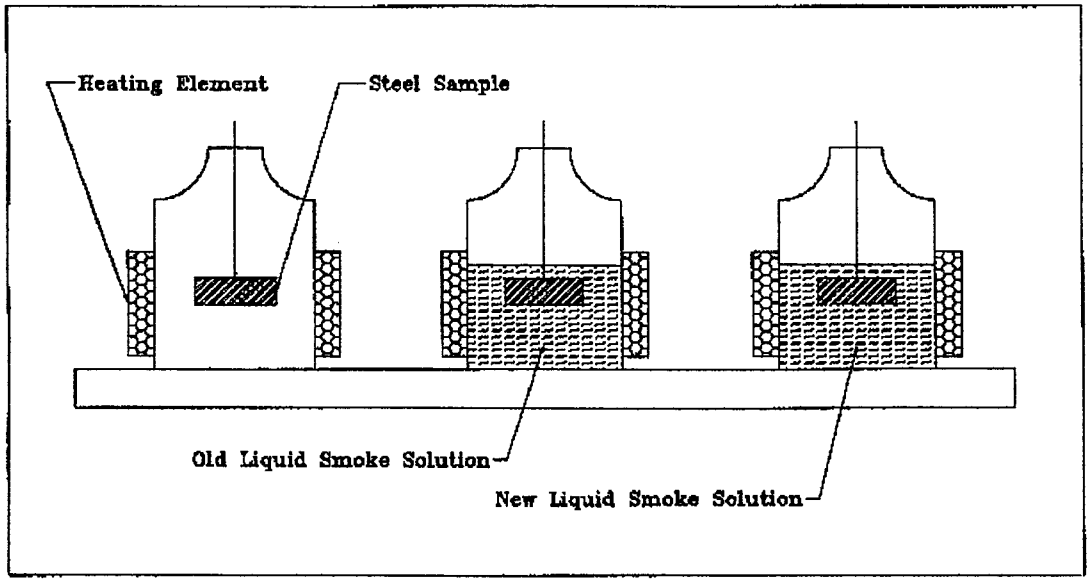


Figure 2: Schematic of the experimental set up for samples dipped in the liquid smoke solutions.

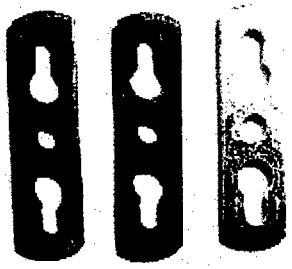


Figure 3: Galvanized Steel Chain Samples, from right to left: not exposed to liquid smoke solution, exposed to old liquid smoke solution and exposed to new liquid smoke solution.

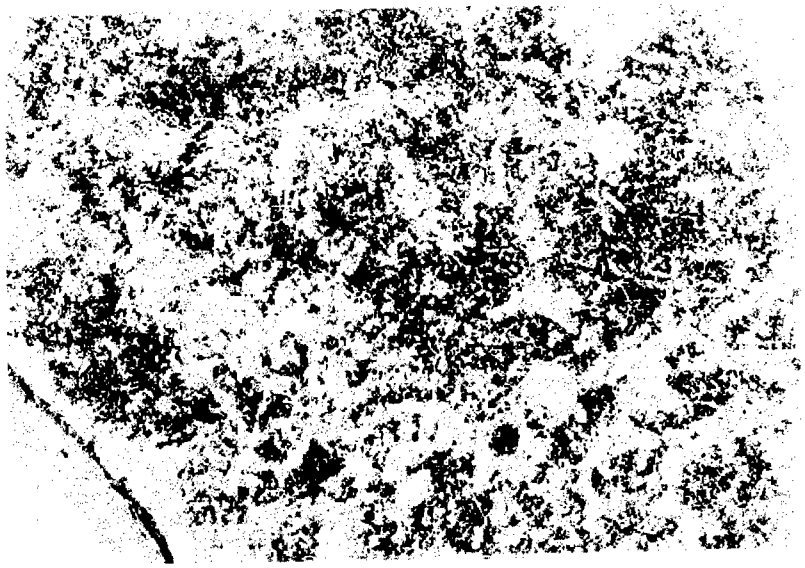


Figure 4: Galvanized Steel Chain Sample, 8x magnification.

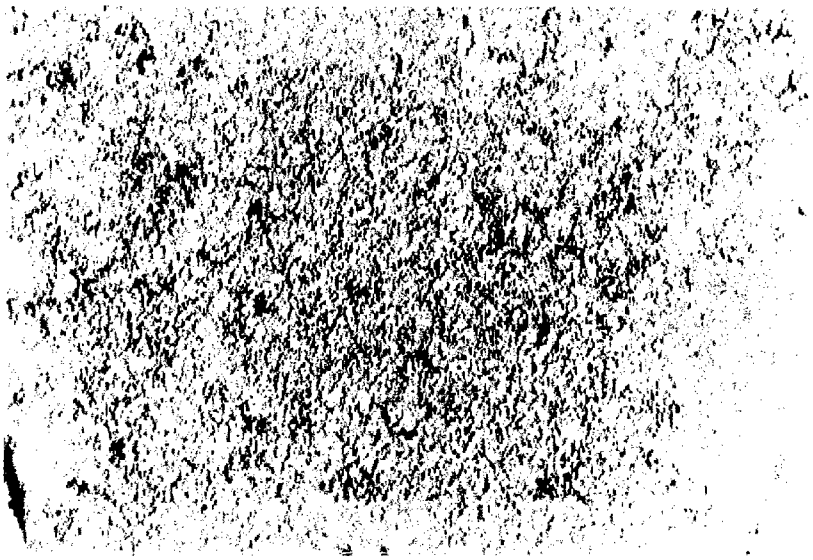


Figure 5: Galvanized Steel Chain Sample exposed to the old liquid smoke solution, 8x magnification.

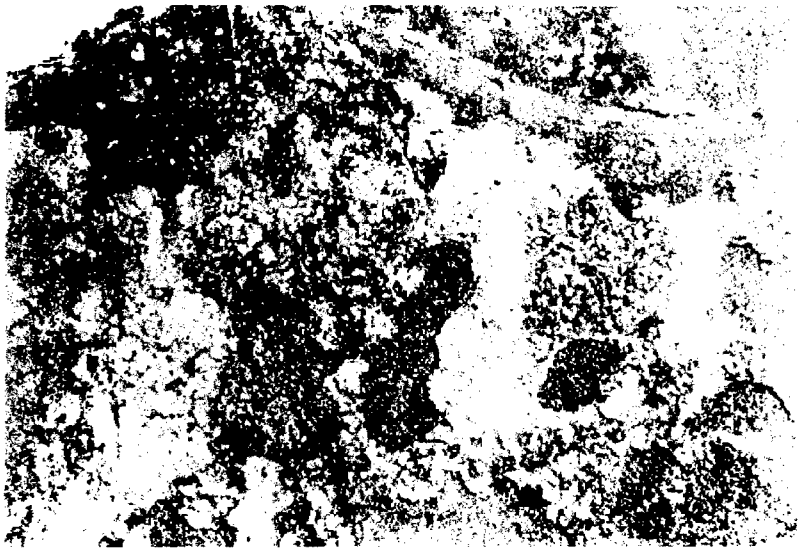


Figure 6: Galvanized Steel Chain Sample exposed to the new liquid smoke solution, 8x magnification.

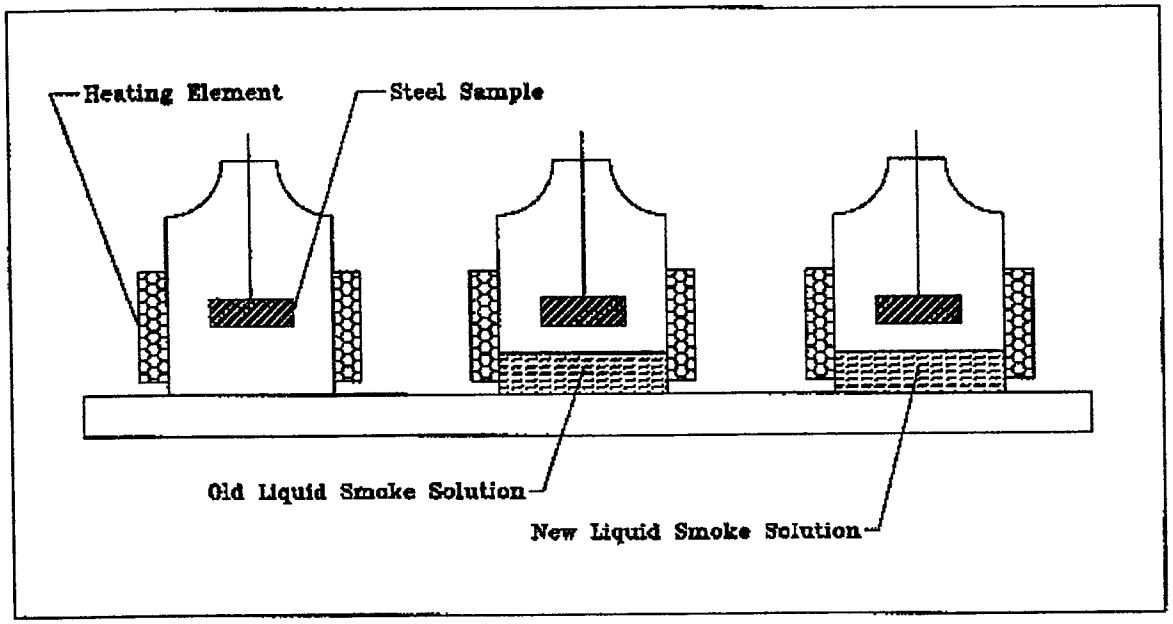


Figure 7: Schematic of the experimental set up for samples placed above the liquid smoke solutions.

OVERALL RANKING:

1) Design Concept A - Change to previous liquid smoke solution.
2) Design Concept B - Change the material to AISI 316.
3) Design Concept C - Add a passivation layer to the AISI 304.

Each of the design solutions offers a viable solution to the problem. Each will solve the problem with respect to the criteria, with only the last option requiring more testing to ascertain feasibility.

CONCLUSION: Extensive testing was done to conclusively determine the cause of the corrosion of the AISI 304 stainless and galvanized steels. Three design concepts were developed that are potential solutions to the current problems at the meat packaging company smokehouse.

REFERENCES

[1] Private Communication with Dr. James Landry, Chemistry Department, Loyola Marymount University, Los Angeles, CA, February 1996.
[2] Tomashow, N.D., The Theory of Corrosion and Protection of Metals, Macmillian, 1966.
[3] Fontana, Mars G., Corrosion Engineering, McGraw Hill, 1986, 361.
[4] Binder, W.O., and Borgmann, C.W., Corrosion of Metals, ASM, 1948.
[5] Monypenny, J.H.G., Stainless Iron and Steel, Chapman & Hall, 1951.
[6] Sedriks, A. John, Corrosion of Stainless Steels, John Wiley & Son, 1979.

STRESS ANALYSIS OF WELDMENTS BY PHOTOSTRESS

Chia-Lung Chang and Ruey-Chung Lee
Department of Mechanical Engineering
National Yunlin Institute of Technology
Toliu, Yunlin, 640 Taiwan

Abstract: The purpose of this research is to study the strain distribution of the weldment under a uniaxially static tension load by the photoelastic method. The shielded metal arc welding (SMAW) was used in this research. Two types of materials, different combinations of weld metal and base metal, are studied. The models were cut from the weldment, perpendicularly to the weld joint. The elastic-plastic behavior of the weld metal, heat affected zone, and base metal can be observed by the fringe pattern of the photoelastic coating. The effects of residual stress on the variation of yielding strength at different locations are also studied. Compared the fringe patterns in models of weldment and plain base metal, the influence of the residual stress on the weldment can be detected. The failure manner of the weldment can be predicted by the principal stress direction on the coating and the location of the densest fringe orders.

Key Words: Photoelastic method; stress analysis; weldment; failure manner

INTRODUCTION: The mechanical characteristics of the weldment are usually determined by the tensile test. While the stress-strain diagram from the test only represents the gross behavior of the weldment, it does not count the material variations among the weld metal, base metal, and heat-affected zone. The chemical compositions and mechanical properties between weld metal and base metal are different. With the appliation of the heat, the mechanical properties of the weldment are changed by the welding. With the applications of strain gage, only localized strain along finite length can be measured. The photostress (photoelastic coating) method can be used to measure the full field strain distributions of the test specimens. Raghavendran and Fourney (1) used two-dimensional photoelastic analysis and two-dimensional finite element method to study the stress distribution across a butt-welded joint. Tsai and Park (2) utilized two-dimensional photoelastic method to determine the stress-intensity-factors of fillet-welded joint. Gambrell (3) and Gambrell and Kavikondala (4) used photoelastic coatings to determine the mechanical properties of aluminum welded joints. Fessler and Pappalettere (5) utilized photoelastic coatings to study the strain distribtuion of fillet-welded joint. In order to measure the full field elastic-plastic strain distribution of the butt-welded steel weldments(weld metal, base metal, and heat-affected zone), photoelastic method was applied to measure the mechanical behaviors in this paper. The correlation between the failure manner in uniaxial tension of the specimen and the fringe pattern observed in the photoelastic coating was also studied.

INSTRUMENTATION AND SPECIMENS: The strain distributions of weldments in tension were measured by the fringe order read in the coatings bonded on the weldments. The relation between the shear strain in the coating and fringe order read can be expressed as follows:

$\varepsilon_x - \varepsilon_y = N \times f$

$\varepsilon_x, \varepsilon_y$: principal strains in the coating

N: fringe order

f: fringe value, obtained from the calibration of coating

Because the weld surface is irregular and contains many grooves, it is difficult to bond the coating. The weld joint was ground off flush with the base metal, then the liquid plastic, PS-8C, contoured on the surface of test part. The material properties of PS-8C are as follows: modulus E-350ksi, maximum elongation 3%. The cement, PC-8, was used to adhere the coating and the test part. The tardy compensation method was used to measure the point-to-point fringe order. All plastic, cement, and photostress equipment were obtained from Measurements Group, Inc., U.S.A. (6)

The welding process of specimens was shielded metal arc welding(SMAW). The specimens were made to satisfy the welding qualifications of ASME requirements. The weldments were made that no defection was found by visual inspection. The welding data of specimens are shown in Table I.

Table I Welding data of specimens

welding process	SMAW	welding operation	manual
electrode	E6010,E7010	SFA No.	SFA 5.1
current type	AC	welding position	1G
welding pass	2	weld joint	double v
current	110A	voltage	220V

1G:1-flat position, G-butt-welded

Two types of material combination were made to measure the strain distributions of specimens : type A weld, the yielding strength of base metal is close to that of electrode; type B weld, the yielding strength of electrode is greater than that of base metal. For type A weld, the base metal used was ASTM A515 carbon steel for pressure vessel plate, and the

electrode used was E6010. Type B weld, the base metal used was JIS SS41 structural steel, the electrode used was E7016. The mechanical properties of the base materials and electrodes used in specimens are showin in Table II.

The specimens were made to satisfy the specification of AWS/AISI(7). The welding width of type A weld was about 10 mm, that of type B weld was about 20 mm. The shape of test specimen is shown in Fig.1. The specimens were cut from a fabricated weldment, 300mm long and 250mm wide, perpendicularly to the weld joint. The ends of the weldment were discarded, because of the non-uniform material properties of the welds during welding.

Table II Standard mechanical properties of the weld

	Yield point strength kg/mm²(N/mm²)	Tesile strength kg/mm²(N/mm²)	Elongation %
ASTM A515	26.5(260) min.	49.5-63.2(485-620)	17
JIS SS41	25 (245)min	41-52(401-509)	17
E6010	34.6(339)	43(421)	22
E7016	41.6(407)	50.5(495)	22

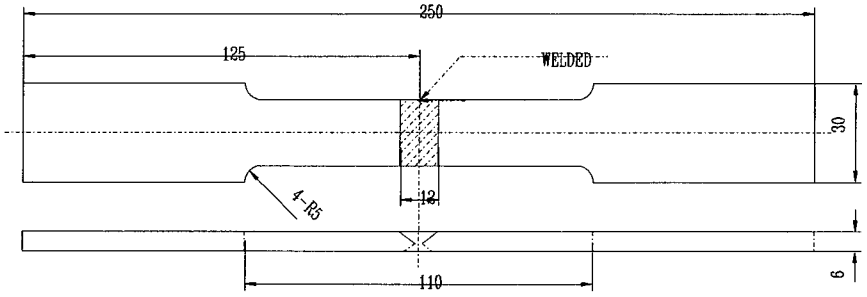


Fig.1 Tensile specimen unit:mm

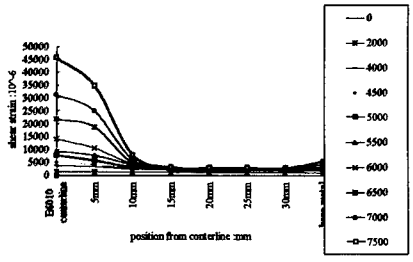


Fig.2 Strain curves of type A weld

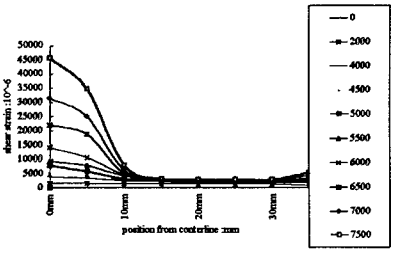


Fig.3 Strain curves of type B weld

RESULTS:

Shear Strain Distributions: The locations of strain measured were taken 5 mm in increments from the centerline of specimens. The strain distributions of type A weld in uniaxial tension are shown in Fig.2. The heat-affected zone (HAZ) is a very thin zone near boundary of base metal and weld metal. It was identified as zone close to the above mentioned boundary. The base metal, HAZ, and weld metal were within elastic region before the load up to 5000kg. When the load increasing gradually above 5000kg, the weld metal exhibited significant plastic strain. The HAZ showed plastic strain when the load above 5500kg. For the base metal near the weld joint (10-15mm) , it did show some plastic strain when the load above 5000kg. But for the base metal far from the centerline (15mm behind), the elastic behavior was not affected by the increasing load. The strain distribution of type B weld in uniaxial tension is shown in Fig.3. The yielding strength of base metal for type B weld is smaller than that of weld metal. The base metal was easier to yield than weld metal. The strain of base metal increased rapidly after the load starting to yield. While the weld metal remained in elastic region.

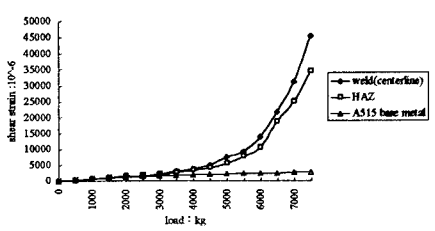


Fig.4 Strain-load curves of type A weld

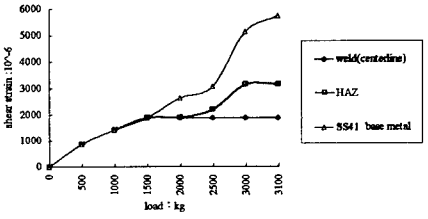


Fig.5 Strain-load curves of type B weld

The strain-load curves of type A weld for weld metal, HAZ, and base metal are shown in Fig.4. The weld metal was identified as the zone at centerline (0 mm), HAZ the zone near 5mm from the centerline, the boundary of weld metal and base metal, base metal the zone 15mm from the centerline, the strain-load curves of various materials were plotted. For type A weld, the weld metal started to yield at load 4000kg (stress=33.33kg/mm^2). From Table II, the yielding strength of the electrode for type A weld, E6010, is 34.6 kg/mm^2. The residual stresses of the weld did have some effects on the yielding strength of the weld metal. The HAZ started to yield at load 4500kg (stress=37.49kg/mm^2). The yielding strength of HAZ is higher than that of base metal. The effect of residual stresses on HAZ was more significant. The base metal was in eleastic regions for all loads. For the type B weld, the zone at centerline (0 mm) was taken as weld metal, the zone near 10 mm from the centerline, the boundary of weld metal and base metal, as HAZ; and the zone 25 mm from the centerline as base metal, the strain-load curves of various materials are shown in Fig.5.

For the type B weld, the yielding strength of base metal was about sixty percents of that of weld metal. The base metal started to yield when the load reached 1500kg. Then the strain increased rapidly. Even the load reaching 3100kg, the weld metal did not exhibit much nonlinearity. The material properties of HAZ are affected by both base metal and weld metal. From the strain-load curves, the mechanical strength of HAZ is between base metal and weld metal. The zone of HAZ was very thin and iregular in shape, only approximate zone could be identified by photostress method. The comparison of strain distributions between weldments and the plain base metal is shown in Fig.6, type A weld vs. A515 base metal; and Fig.7, type B weld vs.SS41 base metal. The effects of residual stresses on the variations in mechanical properties between weldments and the plain base metal were very significant.

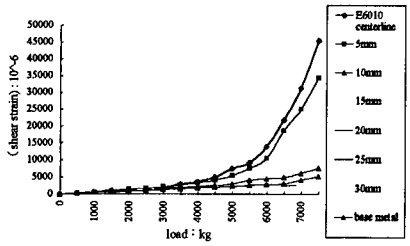


Fig.6 Comparison of strain distributions between type A weld and A515 base metal

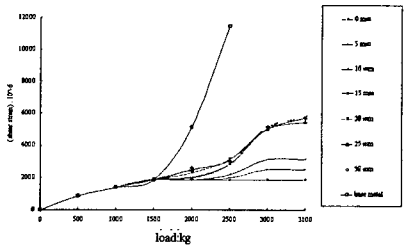


Fig.7 Comparison of strain distributions between type B weld and SS41 base metal

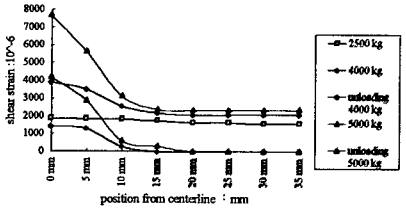


Fig.8 Residual shear strain of type A weld

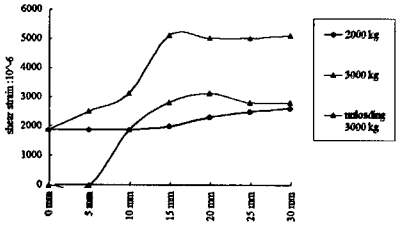


Fig 10 Residual shear strain of type B weld

Fig 9(a) load=5000kg

Fig 9(b) unloading from 5000kg

Fig 9(a) fringe pattern in type A weld

Fig 9(b) fringe pattern of residual strain in type A weld

Residual Shear Strain: The strain left on the coating of weldments after load releasing is called residual strain. The fringe pattern of residual strain for type A weld when unloading from 4000kg is shown in Fig.8. For the weld zone, 0-5mm from centerline, was fully yielded. The zone 5-10mm from the centerline, the material was partially yielded. For the base metal, 15mm from the centerline, it did not exhibit the residual strain. When unloading from 5000kg, the yielded zone increased rapidly, the plastic zone reached near 20 mm from the centerline. The residual strain for type A weld unloading-from 5000kg along the centerline is shown in Fig. 9 (b) The residual strain for type B weld along the centerline is shown in Fig.10. The fringe pattern of residual strain for type B weld when unloading from 3000kg is shown in Fig.11(b) The coating of the weld exhibited more residual strain with unloading from higher load.

Fracture Position: The position of fracture can be predicted by the fringe patter in the coating. The ductility of type A weld was low. The weld position to fracture can be predicted by the starting point of fringe in coating or the position with the densest fringe orders. For type A weld, both points were almost at the same position. The specimen for type A weld after fracture in tension is shown in Fig.12. The elongation of base metal for type B weld was higher than that of the coating. The coating was failed before fracture of the weld. The position to fracture can only be predicted by the starting point of fringe in coating. The specicem for type B weld after fracture in tension is shown in Fig.13. For the observed specimens, the fracture position in tension were quite matched the starting point of fringe in coating and the direction of fracture section was consistent with the isoclinic fringe in coating at the predicted location.

Fig 11(a)
load=3000kg
fringe pattern

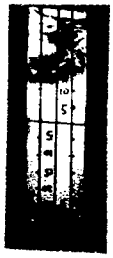


Fig 11(b)
unloading from 3000kg
residual strain

Fig 11(a) fringe pattern in type B weld
Fig 11(b) fringe pattern of residual strain in type B

Fig 12 type A specimen after fracture

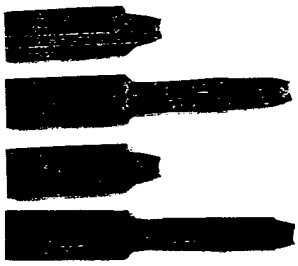


Fig 13 type B specimen after fracture

CONCLUSIONS: The applications of photoelastic method to measure the point-to-point mechanical properties of weldments is very useful. It does not average the strain in finite length like strain gages and shows the strain distributions in full field for various parts of the weld. Compared the fringe patterns of weldments and plain base metal, the effects of welding process on the welds can easily be dectected. The positions of fracture in tension for two types of specimen can be predicted by the starting point of fringe in coating.

ACKNOWLEDGEMENT: This work was supported in part by the National Science Council of R.O.C. under contract no. NSC-85-2211-E-224-002. The authors would like to express their thanks herewith.

REFERENCES:

1. N.S. Raghavendran and M.E. Fourney , 1994, "Stress Analysis of a Welded Joint," Engineering Fracture Mechanics, V48, N5, pp.619-627.
2. C.L. Tsai and S.K. Park, September 1984, "Determination of Stress-Intensity Factors of Fillet-Welded T-Joint by Computer-Assisted Photoelasticity, " Experimental Mechanics, pp.233-242.
3. S.C. Gambrell, Jr. , January/February 1993, "Use of Photostress to Characterize the Mechanical Behavior of Weldments, " Experimental Techniques, pp.15-18.
4. S.C. Gambrell, Jr. and K. Kavikondala, March/April 1994, "Yield Detection in Aluminum Welded Joints Using Photostress, " Experimental Techniques, pp.11-14.
5. H. Fessler and C. Pappalettere , 1989, "Plastic-Elastic Strains in Two-Dimensional Sections of Partial-Penetration Fillet Welds" , Journal of Strain Analysis, V 24, N 1, pp15-21.
6. Measurement Group Inc. , 1981, Instroduction to Stress Analysis by the Photo Stress Method, TN-702, U.S.A.
7. ASME, 1992, "ASME Boiler And Pressure Vessel Code, " SECTION II PART C Specifications for Welding Rods,Electrodes, and Filler Metal , pp1-36.8, U.S.A.

INTERLAYER SHEAR SLIP EFFECTS ON THE LOAD TRANSFER MECHANISM IN NONSYMMETRICALLY CONSTRAINED THICK LAMINATED STRIPS

Dwight W. Swett
Hughes Aircraft Company
Mail Station EO/E1/D102
El Segundo, CA 90245
(310) 616-1335

Abstract: In this paper, the plane elasticity solution for the stresses and displacements in a multilayer laminated strip subjected to axially symmetric transverse load is developed including the edge effects due to end condition. The solution corresponds to the shear free end condition neglected in previous works based on enforced transverse displacement conditions at the strip ends. This feature of the solution allows the effects of transversely nonsymmetric end constraint on interlaminar shear deformation to be included in the analysis. The resulting eigenfunctions satisfy the necessary equations of equilibrium, conditions of displacement compatibility, requirements on interlaminar stress and displacement continuity, and shear traction free edge conditions. Convergence for enforced zero transverse slope at the strip ends can be established, as the eigenfunctions are orthogonal. Consequently, the solution is exact for the edge conditions corresponding to shear free guided ends. Numerical results are presented for several examples to examine the effects of transversely nonsymmetric end condition on interlaminar shear growth and shear transfer in the presence of interlayer shear slip. The interlaminar shear distributions indicate the impacts of the edge effect are particularly significant in thick laminates and are shown to be the dominant mechanism controlling shear transfer between the layers once interlayer shear slip has developed. In fact, these results indicate load transfer can be severely limited in thick laminates having transversely nonsymmetric boundary conditions when even moderate interlayer shear slip occurs. For certain laminate configurations, a radical departure from established solutions which neglect the edge effects can be seen.

Key Words: Edge Effects; Interlayer Shear Slip; Shear Deformation; Shear Transfer; Thick Laminate; Transverse Loading

1 Introduction

One of the most severe problems associated with the use of laminated structures in mechanical loading environments is the susceptibility to failure due to imperfect interfacial bonding resulting from bond joint compliance or loss of bond joint integrity. The effects of interfacial integrity on load transfer between the constituent laminae are well documented and can be the primary mechanism behind premature failure of these types of structures under mechanical loading. In addition, the problem may be compounded by the introduction of lateral surface supporting mechanisms which do not uniformly distribute loads across a section of the laminate. Trade studies to develop a satisfactory design for these types of mechanical environments have heretofore been rather limited due to the lack of accurate

analytical assessments for these interface and edge effects that can be catastrophic to the structure. The predominant amount of investigation for these types of mechanical loading problems has been restricted to detailed numerical finite element analyses that do not lend to rapid concurrent engineering design processes. No analytical solution has been available to address the combined factors of edge effects and interfacial influence in laminates subjected to mechanical loads.

Historically, the problem of stresses in mechanically loaded laminates has been addressed by formulating the edge conditions in terms of equipollent forces and moments with reliance on Saint-Venant's Principle to ensure satisfactory stress distributions away from the edges. This is the motivation for the thin assumption invoked in classical laminate theory. To address this difference in laminates which are not sufficiently thin enough to invoke this assumption, many investigators have included transverse shear deformation effects in an attempt to account for the stress distribution in the vicinity of the edges. One of the earliest of these is credited to Yu [19] and was based upon a piecewise linear assumption for displacement variation through the thickness. Other approximations by Yang *et. al.* [18], Whitney and Pagano [15], Reissner [9], and Whitney and Sun [17], to mention some, utilize variations on this same displacement approach. These types of "shear deformation" theories lead to accurate deflections in bending problems but have been shown by Whitney [16] to provide no improvement in the stress field compared to classical plate theory near the boundary layer. Though each of these is different in the choice of displacement distribution, all have discontinuous tractions across the laminae interfaces. Pagano [6] developed an exact elasticity solution to the problem of composite laminates in cylindrical bending, but the boundary conditions included a specified transverse displacement condition which eliminated the incorporation of any transversely nonsymmetric edge effects. Later, Pagano [7] applied the variational theorem of Reissner [8] to address the free edge effects in the end problem of an infinitely long laminate by assuming linear stress variations through each layer. The solution satisfies interlaminar stress and displacement continuity requirements but addresses the end conditions on a "layer equilibrium" basis only. Hess [4] developed an elasticity solution to the edge effects for the end problem using a series of nonorthogonal Fadle-Papkovitsch eigenfunctions. The eigenfunctions are each a solution to the differential equations of compatibility and satisfy the traction-free boundary conditions on the lateral surfaces as well as interfacial continuity requirements. The solution is not exact as the end traction conditions are approximated using a least squares fit to the distribution. While approximations to the classical end problem having shear free edges are available, a lack of analytical solutions exists to address the edge effects in transverse load problems.

In addition to the influence of edge condition on the structural performance of thick laminates, it is widely recognized that the interfacial condition plays a significant role in the load capability of the structure. Many investigators have developed approximate solutions accounting for nonrigid interfaces in laminated structures. Most of these are based upon a linear shear slip law applied to classical Bernoulli-Euler beam theory and do not include shear deformation effects. Among these are included the work of Goodman and Popkov [2], Goodman [3], Thompson *et al.* [12], and Vanderbilt *et al.* [14]. Toledano and Murakami [13] applied a linear shear slip law to a laminate theory accounting for interlaminar shear deformation and interfacial shear stress continuity. However, the interlaminar shear stresses that governed the shear slip at the interface were not accurately predicted by the theory due to the low order of the assumed displacement field. Lu and Liu [5] later incorporated a linear shear slip law into a higher order shear deformation theory for laminated beams that satisfies interlaminar stress continuity.

In this paper, a solution for the deformation and stress state in a multilayer strip subjected to an arbitrary axially symmetric transverse surface load is developed. A departure from previous analyses involving a series expansion in the thickness coordinate resulting in a non orthogonal Fadle-Papkovitsch eigenfunction was taken. The present solution assumes a length coordinate expansion of the stresses and displacements, providing an exponential variation in the thickness coordinate direction. This results in an orthogonal series representation suitable for modeling arbitrarily symmetric transverse loads as well as the local end conditions. In addition, the approach allows satisfaction of all edge conditions solely with the eigenvalues and provides a system of $4k$ equations linear in the $4k$ coefficients necessary to satisfy the interface continuity requirements, k being the number of layers in the lamination. For a truncated finite series approximation of N terms, a set of N such linear systems must be solved for the entire solution. The displacement compatibility requirement, interlaminar displacement and stress continuity requirements, and shear traction free edge boundary conditions are then satisfied exactly. The edge traction and geometric conditions satisfied correspond to those of enforced zero transverse slope ($\frac{\partial w}{\partial x} = 0$) at the strip ends. The equilibrium requirement is satisfied to the exact extent that the finite Fourier series representation converges to the applied surface traction distributions. Thus, the solution is exact within plane elasticity theory for shear free guided ends. Bonding layers or "interlayer slip" zones can then be accounted for directly owing to the exact nature of the solution.

2 Description of the Mathematical Model

Consider an arbitrary multilayer laminated strip formed by creating a perfect bond between a number of rectangular, homogeneous, elastic strips of equal length l. It is desired to determine the stresses and deformations in the laminate for the case in which a transverse surface traction and equilibrating reaction are imposed on the structure while the ends remain free of shear traction. It is assumed that the applied surface traction is distributed

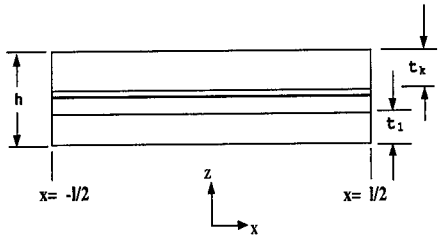


Figure 1: Laminated Strip Geometry

symmetrically along the strip direction. The geometry considered is shown in Figure 1. In addition, it is required that the slope of the transverse displacement $\frac{\partial w}{\partial x}$ at the ends of the strip remain zero.

3 Analysis

3.1 Differential Equations of Compatibility and Equilibrium

In the classical theory of elasticity, we are concerned with the determination of a stress or strain state in the interior of a body subject to a given state of stress or strain (or displacement) on the boundary of the body. The transversely loaded multilayer strip of different materials falls in this category due to the interlaminar shear transfer between the layers and the continuous displacement field through the laminate. Here we develop the governing differential equations for the complete solution of the generalized plane stress case.

In the tensor notation, the following equilibrium equations:

$$\frac{\partial}{\partial x_j}\sigma_{ij} + \rho\beta_i = \rho\frac{\partial^2}{\partial t^2}u_i \tag{1}$$

and strain displacement relations :

$$e_{ij} = \frac{1}{2}(\frac{\partial}{\partial x_j}u_i + \frac{\partial}{\partial x_i}u_j) \tag{2}$$

remain valid for the plane stress problem. For isotropic materials, the constitutive relations for small strains can be described as follows:

$$\sigma_{ij} = \lambda e_{kk}\delta_{ij} + 2\mu e_{ij} \tag{3}$$

where e_{kk} is the bulk dilatation, and λ and μ are the Lamé constants. For the homogeneous part of the solution, we reconsider the equilibrium requirement in Eq. 1 in the absence of body forces, dynamic effects, and temperature loading in unabridged notation:

$$\frac{\partial}{\partial x}\sigma_{xx} + \frac{\partial}{\partial z}\sigma_{xz} = 0 \tag{4}$$

$$\frac{\partial}{\partial x}\sigma_{xz} + \frac{\partial}{\partial z}\sigma_{zz} = 0 \tag{5}$$

which suggests equilibrium can be satisfied with an Airy stress function solution ψ defined by:

$$\sigma_{xx} = \frac{\partial^2}{\partial z^2}\psi \tag{6}$$

$$\sigma_{zz} = \frac{\partial^2}{\partial x^2}\psi \tag{7}$$

$$\sigma_{xz} = -\frac{\partial^2}{\partial x \partial z}\psi \tag{8}$$

To ensure compatibility of the displacements associated with the stress function, we require:

$$\frac{\partial^2}{\partial z^2}e_{xx} + \frac{\partial^2}{\partial x^2}e_{zz} = 2\frac{\partial^2}{\partial x \partial z}e_{xz} \tag{9}$$

which, upon substitution of Eq. 6 through Eq. 8 into the constitutive relations of Eq. 3 and evaluating for strain components e_{ij}, we obtain in terms of the Airy stress function ψ, the necessary requirement for the homogeneous part of the stress solution:

$$\nabla^4\psi = 0 \tag{10}$$

3.2 Eigenfunction Solutions

The general solution for the homogeneous differential shown in Eq. 10 can be determined from a separation of variables approach to be:

$$\begin{aligned}\psi &= [(a_1 + a_5 z)\cosh(\lambda z) + (a_2 + a_6 z)\sinh(\lambda z)]\cos(\lambda x) + \\ &\quad [(a_3 + a_7 z)\cosh(\lambda z) + (a_4 + a_8 z)\sinh(\lambda z))\sin(\lambda x) + \\ &\quad [(a_9 + a_{13} x)\cosh(\beta x) + (a_{10} + a_{14} x)\sinh(\beta x))\cos(\beta z) + \\ &\quad [(a_{11} + a_{15} x)\cosh(\beta x) + (a_{12} + a_{16} x)\sinh(\beta x))\sin(\beta z) + \\ &\quad \sum_{m=0}^{1}\sum_{n=0}^{1}(A_{mn} x^{m+2} z^n + B_{mn} x^m z^{n+2} + C_{mn} x^m z^n) \end{aligned} \tag{11}$$

For symmetric boundary conditions at the strip ends, however, we realize all nonsymmetric terms in x are necessarily zero. Therefore, we can reduce the stress function relation of Eq. 11 to:

$$\begin{aligned}\psi &= [(a_1 + a_5 z)\cosh(\lambda z) + (a_2 + a_6 z)\sinh(\lambda z)]\cos(\lambda x) + \\ &\quad [(a_9 \cosh(\beta x) + a_{14} x \sinh(\beta x)]\cos(\beta z) + \\ &\quad [(a_{11}\cosh(\beta x) + a_{16} x \sinh(\beta x)]\sin(\beta z) + \\ &\quad \sum_{n=0}^{1}(A_{0n} x^2 z^n + B_{0n} z^{n+2} + C_{0n} z^n) \end{aligned} \tag{12}$$

Now, two methods of expansion seem available, one in the thickness coordinate z, or the other along the strip in x. The conventional approaches have taken an expansion in the thickness coordinate resulting in a set of non orthogonal Fadle-Papkovitsch eigenfunctions. This approach results in an exponential variation of the solution along the strip direction and a set of non orthogonal eigenfunctions in the thickness coordinate that restrict this approach to those problems involving mechanical end loads or uniform temperature distributions. Since we are interested in nonuniform transverse load effects, here we take the latter approach and expand the solution as a function of the axial strip coordinate. Imposing the shear traction-free edge requirement, we have from the shear stress function relation of Eq. 8:

$$\sin(\lambda \frac{l}{2}) = 0 \tag{13}$$

or:

$$\lambda = \lambda_n = \frac{2n\pi}{l} \tag{14}$$

providing, after appropriate elimination of non orthogonal series terms:

$$\begin{aligned}\psi(x,z) &= \sum_{n=1}^{\infty}\{[(A_n + C_n z)\cosh(\lambda_n z) + (B_n + D_n z)\sinh(\lambda_n z)]\}\cos(\lambda_n x) + \\ &\quad F_0 x^2 + F_2 z^2 + F_3 z^3 \end{aligned} \tag{15}$$

where terms in $(x^2 z)$ have been eliminated for satisfaction of the shear traction free edge conditions and all polynomial terms up to first order, being eliminated in the stress relations, are arbitrary and set to zero.

The stresses resulting from Eq. 15 are formed from a symmetric set of orthogonal eigenfunctions of the axial coordinate. Consequently, any (symmetric) lateral surface traction

can be represented completely by the series of eigenfunctions to within a Fourier series constant. As a result all polynomial terms in Eq. 15 not related to transverse stress will be necessarily zero. For the resulting stresses in the layer, we can obtain with the appropriate substitutions and operations:

$$\sigma_{xx} = \sum_{n=1}^{\infty} \left\{ \frac{2n\pi}{l} \left[\left(\frac{2n\pi}{l}(A_n + C_n z) + 2D_n \right) \cosh\left(\frac{2n\pi}{l} z\right) + \left(\frac{2n\pi}{l}(B_n + D_n z) + 2C_n \right) \sinh\left(\frac{2n\pi}{l} z\right) \right] \right\} \cos\left(\frac{2n\pi}{l} x\right) \quad (16)$$

$$\sigma_{zz} = 2\,F_0 - \sum_{n=1}^{\infty} \left\{ \left[\left(\frac{2n\pi}{l}\right)^2 \left((A_n + C_n z) \cosh\left(\frac{2n\pi}{l} z\right) + (B_n + D_n z) \sinh\left(\frac{2n\pi}{l} z\right) \right) \right] \right\} \cos\left(\frac{2n\pi}{l} x\right) \quad (17)$$

$$\sigma_{xz} = \sum_{n=1}^{\infty} \frac{2n\pi}{l} \left[C_n \cosh\left(\frac{2n\pi}{l} z\right) + D_n \sinh\left(\frac{2n\pi}{l} z\right) \right] \sin\left(\frac{2n\pi}{l} x\right) + \sum_{n=1}^{\infty} \left(\frac{2n\pi}{l}\right)^2 \left[(A_n + C_n z) \sinh\left(\frac{2n\pi}{l} z\right) + (B_n + D_n z) \cosh\left(\frac{2n\pi}{l} z\right) \right] \sin\left(\frac{2n\pi}{l} x\right) \quad (18)$$

and, for associated displacements, we can obtain (after a lengthy calculation):

$$u = \frac{1}{E} \sum_{n=1}^{\infty} \left[\left((1+\nu) \frac{2n\pi}{l}(A_n + C_n z) + 2D_n \right) \cosh\left(\frac{2n\pi}{l} z\right) \right] \sin\left(\frac{2n\pi}{l} x\right) + \frac{1}{E} \sum_{n=1}^{\infty} \left[\left((1+\nu) \frac{2n\pi}{l}(B_n + D_n z) + 2C_n \right) \sinh\left(\frac{2n\pi}{l} z\right) \right] \sin\left(\frac{2n\pi}{l} x\right) \quad (19)$$

$$w = w_0 + 2\,\frac{1}{E} F_0\, z - \frac{1}{E} \sum_{n=1}^{\infty} \left[\left((1+\nu) \frac{2n\pi}{l}(A_n + C_n z) - (1-\nu) D_n \right) \sinh\left(\frac{2n\pi}{l} z\right) \right] \cos\left(\frac{2n\pi}{l} x\right) - \frac{1}{E} \sum_{n=1}^{\infty} \left[\left((1+\nu) \frac{2n\pi}{l}(B_n + D_n z) - (1-\nu) C_n \right) \cosh\left(\frac{2n\pi}{l} z\right) \right] \cos\left(\frac{2n\pi}{l} x\right) \quad (20)$$

The stresses and deformations of Eq. 16 through Eq. 20 comprise the exact solution for the plane elasticity problem having enforced zero transverse slope end conditions and shear traction free edges.

4 Transverse Loading

For the general laminate having an axially symmetric upper lateral surface traction $P(x)$ representable in Fourier series form as:

$$P(x) = P_o + \sum_{n=1}^{\infty} P_n \cos\left(\frac{2n\pi x}{l}\right) \quad (21)$$

and having statically equivalent supports on either lateral surface locally distributed over the region $\xi_1 < |x| < \xi_2$ with uniform pressure, we realize that for any symmetrically distributed transverse loading the support reaction can be represented in the form:

$$R(x) = \pm \left[P_o + \sum_{n=1}^{\infty} P_o \left\{ \frac{l}{n\pi(\xi_2 - \xi_1)} \left(\sin\left(\frac{2n\pi\xi_2}{l}\right) - \sin\left(\frac{2n\pi\xi_1}{l}\right) \right) \right\} \cos\left(\frac{2n\pi x}{l}\right) \right] \quad (22)$$

where positive sign applies for lower lateral surface reactions and negative sign applies to upper lateral surface reactions.

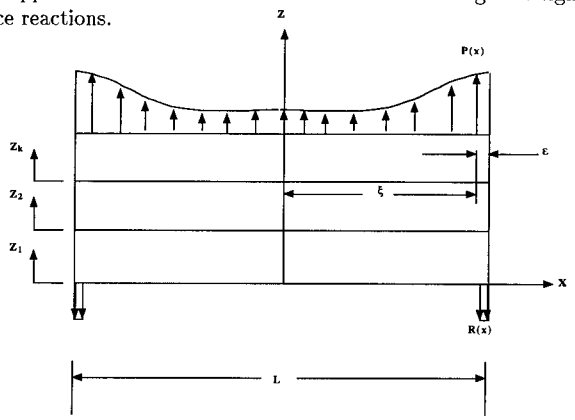


Figure 2: Symmetric Transverse Load Distribution.

To couple in the edge effects due to the transverse load distribution, an approach similar to that utilized by Swett and Shiflett [10], [11] for temperature loading is utilized here as well where the finite series representation of Eq. 21 must be taken in the limit as ξ approaches $\frac{l}{2}$ to include these terms in the homogeneous solution. Consequently, the associated Fourier series coefficients are given by the relations:

$$P_0 = \frac{1}{l}\left(\int_{-\xi}^{\xi} P(x)\, dx\right) \quad (23)$$

$$P_n = \frac{2}{l}\left(\int_{-\xi}^{\xi} P(x) \cos\left(\frac{2n\pi x}{l}\right) dx\right) \quad (24)$$

where $\xi = (\frac{l}{2} - \epsilon)$ is the analysis region for the limit as ϵ approaches zero as shown in Figure 2 and *P(x)* is the axially symmetric transverse load.

5 Multilayer Laminate

In this section we utilize the preceding stress and displacement results to develop the linear system of equations for a multilayered laminate of arbitrary number of layers k subjected to an axially symmetric transverse load distribution. The equations arise from the remaining interface continuity and loaded lateral surface boundary requirements. As an illustration of the method of application, we consider the case of transversely loaded upper lateral surface with statically equivalent lower surface reaction.

For the general laminate having transversely loaded lateral surfaces, we require equlibrium of surface normal stresses satisfying the boundary equlibrium requirement:

$$\sigma_{ij}\, n_j = T_i \tag{25}$$

where n_j is the j^{th} component of the outward directed surface normal vector and T_i is the component of the first order surface traction tensor parallel to the i^{th} coordinate axis.

For shear free lateral surfaces, letting the superscript $^{(i)}$ denote the coefficient of the ith laminae, we then have at the lower laminate surface:

$$\begin{aligned} \sigma_{zz}^{(1)}|_{z_1=0} &= R(x) \\ A_n^{(1)} &= P_o\left\{\frac{l}{n\pi(\xi_2-\xi_1)}\left(\sin\left(\frac{2n\pi\xi_2}{l}\right)-\sin\left(\frac{2n\pi\xi_1}{l}\right)\right)\right\} \qquad (26) \\ \sigma_{xz}^{(1)}|_{z_1=0} &= 0 \\ \frac{2n\pi}{l}B_n^{(1)}+C_n^{(1)} &= 0 \qquad (27) \end{aligned}$$

Similarly at the upper laminate surface:

$$\begin{aligned} \sigma_{zz}^{(k)}|_{z_k=t_k} &= P(x) \\ (A_n^{(k)}+C_n^{(k)}t_k)\cosh\left(\frac{2n\pi}{l}t_k\right)+(B_n^{(k)}+D_n^{(k)}t_k)\sinh\left(\frac{2n\pi}{l}t_k\right) &= P_n \qquad (28) \\ \sigma_{xz}^{(k)}|_{z_k=t_k} &= 0 \\ \left(D_n^{(k)}+\frac{2n\pi}{l}(A_n^{(k)}+C_n^{(k)}t_k)\right)\sinh\left(\frac{2n\pi}{l}t_k\right)+ & \\ \left(C_n^{(k)}+\frac{2n\pi}{l}(B_n^{(k)}+D_n^{(k)}t_k)\right)\cosh\left(\frac{2n\pi}{l}t_k\right) &= 0 \qquad (29) \end{aligned}$$

At the laminae interfaces, we require continuity of peel and shear stresses for equilibrium. Therefore:

$$\begin{aligned} \sigma_{zz}^{(i)}|_{z_i=t_i} &= \sigma_{zz}^{(i+1)}|_{z_{i+1}=0} \\ A_n^{(i+1)}-\Big[(A_n^{(i)}+C_n^{(i)}t_i)\cosh\left(\frac{2n\pi}{l}t_i\right)+ & \\ (B_n^{(i)}+D_n^{(i)}t_i)\sinh\left(\frac{2n\pi}{l}t_i\right)\Big] &= 0 \qquad (30) \\ \sigma_{xz}^{(i)}|_{z_i=t_i} &= \sigma_{xz}^{(i+1)}|_{z_{i+1}=0} \\ \frac{2n\pi}{l}\Big[(A_n^{(i)}+C_n^{(i)}t_i)\sinh\left(\frac{2n\pi}{l}t_i\right)+ & \\ (B_n^{(i)}+D_n^{(i)}t_i)\cosh\left(\frac{2n\pi}{l}t_i\right)\Big]+ & \\ C_n^{(i)}\cosh\left(\frac{2n\pi}{l}t_i\right)+D_n^{(i)}\sinh\left(\frac{2n\pi}{l}t_i\right)- & \\ \left(\frac{2n\pi}{l}B_n^{(i+1)}+C_n^{(i+1)}\right) &= 0 \qquad (31) \end{aligned}$$

For interface displacement continuity:

$$u^{(i)}|_{z_i=t_i} = u^{(i+1)}|_{z_{i+1}=0}$$

$$\frac{1}{E_i}\left((1+\nu_i)\frac{2n\pi}{l}(A_n^{(i)}+C_n^{(i)}t_i)+2D_n^{(i)}\right)\cosh\left(\frac{2n\pi}{l}t_i\right)+$$
$$\frac{1}{E_i}\left((1+\nu_i)\frac{2n\pi}{l}(B_n^{(i)}+D_n^{(i)}t_i)+2C_n^{(i)}\right)\sinh\left(\frac{2n\pi}{l}t_i\right)-$$
$$\frac{1}{E_{i+1}}\left((1+\nu_{i+1})\frac{2n\pi}{l}A_n^{(i+1)}+2D_n^{(i+1)}\right) = 0 \qquad (32)$$

$$w^{(i)}|_{z_i=t_i} = w^{(i+1)}|_{z_{i+1}=0}$$

$$\frac{1}{E_i}\left((1+\nu_i)\frac{2n\pi}{l}(A_n^{(i)}+C_n^{(i)}t_i)-(1-\nu_i)D_n^{(i)}\right)\sinh\left(\frac{2n\pi}{l}t_i\right)+$$
$$\frac{1}{E_i}\left((1+\nu_i)\frac{2n\pi}{l}(B_n^{(i)}+D_n^{(i)}t_i)-(1-\nu_i)C_n^{(i)}\right)\cosh\left(\frac{2n\pi}{l}t_i\right)-$$
$$\frac{1}{E_{i+1}}\left((1+\nu_{i+1})\frac{2n\pi}{l}B_n^{(i+1)}-(1-\nu_{i+1})C_n^{(i+1)}\right) = 0 \qquad (33)$$

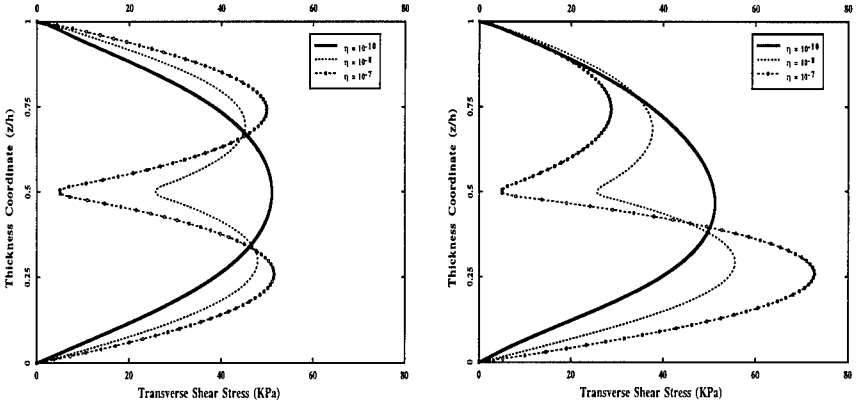
With $(k-1)$ laminae interfaces governed by Eqs. 30 through 33 and two free lateral surfaces governed by Eqs. 26 and 27 (bottom surface) and Eqs. 28 and 29 (top surface), we now have a system of $4k$ equations in the $4k$ unknowns A_n, B_n, C_n, and D_n for the laminate stresses σ_{zz}, σ_{xz}, and σ_{xx}.

This completely defines the exact solution for the plane elastic problem of transverse loading on a laminated strip having enforced zero transverse slope at the strip ends and shear traction free edges. Quantitative results are presented in the following section.

6 Numerical Results

6.1 Symmetric Three Layer Strip: Concentrated Center Load

To investigate the effects of lateral-surface constraint on the stress distribution in thick laminated strips, here we consider a strictly symmetrical problem involving two layers comprised of the same material joined with a thin compliant joint. As material dissimilarity



(a) (b)

Figure 3: Interlaminar Shear Stress Distributions Through the Thickness of the Symmetric Strip at $x = -\frac{3}{8}\,l$, S=4: (a) Symmetric End Condition, (b) Nonsymmetric End Condition

will affect load distribution, the two layers are assumed of the same material for this example to gain an accurate assessment of the effects of interlayer shear slip and edge condition on load distribution. A range of joint compliance is utilized to assess the trends and impacts of interlayer shear slip on practical design and analysis. We seek to determine the deformations and stress distribution due to a concentrated center load of the form:

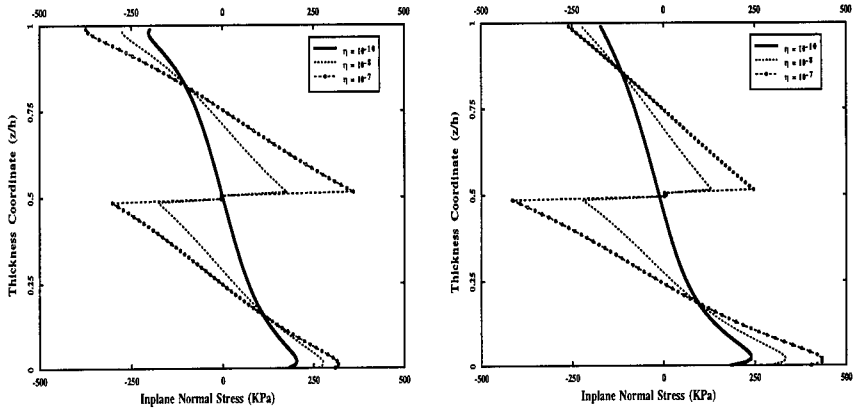
$$P(x) = \begin{cases} 0 & (|x| > \eta) \\ P_{max} & (|x| \leq \eta) \end{cases} \tag{34}$$

where P_{max} corresponds to a 0.69 MPa (100 psi) pressure load applied over $(|x| \leq \eta)$. Reaction load is imposed as local transverse pressure at the ends of the strip on upper or lower surfaces of the form:

$$R(x) = \begin{cases} 0 & (|x| < \xi_1) \\ \pm P_{max} \left(\frac{\eta}{\xi_2 - \xi_1}\right) & (\xi_1 \leq |x| \leq \xi_2) \end{cases} \tag{35}$$

where $\xi_1 = 0.90\frac{l}{2}$, $\xi_2 = 0.993\frac{l}{2}$, and $\eta = 0.025\frac{l}{2}$. The bonded strip is 10 mm (0.4 inch) in length and consists of two 1.25 mm (0.05 inch) layers joined with a 0.05 mm (0.002 inch) interface joint. For strictly computational purposes, we utilize the following isotropic properties at the two layers:

$$E_{layer} = 172.3\,GPa\ (25\,MPsi)$$
$$\nu_{layer} = 0.45$$



(a) (b)

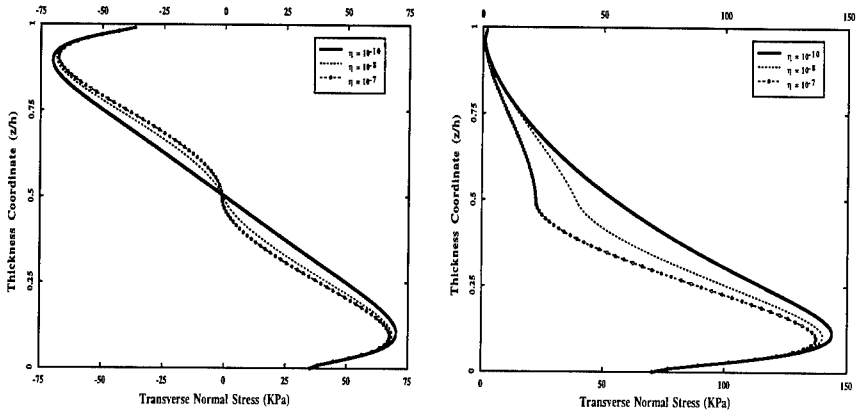
Figure 4: Inplane Normal Stress Distributions Through the Thickness of the Symmetric Strip at $x = -\frac{1}{2}\,l$, S=4: (a) Symmetric End Condition, (b) Nonsymmetric End Condition

The three layer strip described was analyzed by the analytical method described earlier in this paper. The nonsymmetric end condition with transverse reactions taken out on the lower lateral surface is compared to a strictly symmetric end condition having equivalent reactions taken out at both the upper and lower surfaces. A range of interface bond compliance is considered to evaluate the effects of interlayer shear slip on load redistribution. The results are shown in the form of a series of plots for stress distribution as a function of thickness coordinate (z/h). The interlaminar shear and transverse normal stress results of the present solution are shown in Figure 3 and Figure 5 for a range of interface bond compliance of ($10^{-10} \leq \eta \leq 10^{-7}$) where η denotes a measure of the relative interface bond compliance given by:

$$\eta = \frac{t_{bond}}{G_{bond}}$$

The inplane normal stress results of the present solution are shown in Figure 4 for the same range of interface bond compliances. The analytical results were calculated using FORTRAN with a finite series approximation of $N = 45$ terms required for convergence.

Observation of the results for interlaminar shear distributions in Figure 3a indicates that for symmetric boundary conditions, the shear distribution remains symmetrical with both layers sharing shear load equally over the range of interlayer shear slip. The bending stress distributions of Figure 4a for symmetric boundary conditions indicate a slightly larger portion of the bending load taken by the upper layer as interlayer shear slip develops. For large interface compliance ($\eta = 10^{-7}$), about 25% larger peak bending stress is indicated in the top layer than the bottom layer. As the applied load remains the only nonsymmetric factor in the problem, this difference in bending load distribution between the two layers can be attributed to loading the upper surface with the concentrated center load in the presence of guided ends. Comparison of the nonsymmetric end condition results for interlaminar shear distributions in Figure 3b indicates virtually no difference in load distribution when compared with the symmetric boundary condition for rigid interfaces. However, a marked



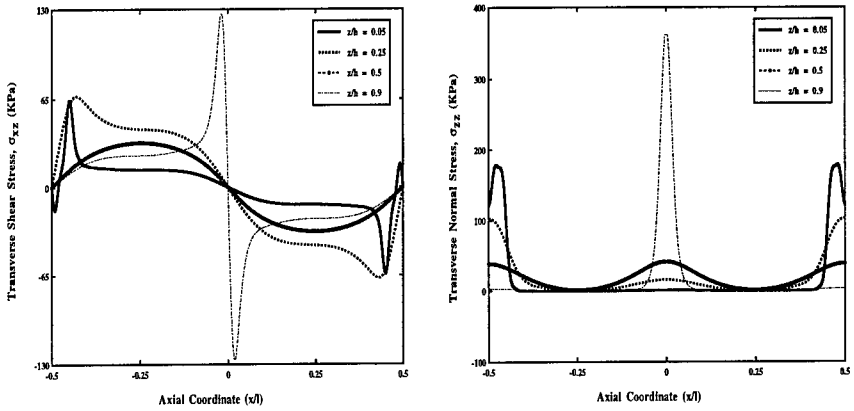
(a) (b)

Figure 5: Transverse Normal Stress Distributions Through the Thickness of the Symmetric Strip at $x = -\frac{1}{2}\,l$, S=4: (a) Symmetric End Condition, (b) Nonsymmetric End Condition

difference in the load distribution is observed between the two boundary conditions as shear slip develops at the interface. The results for the lower surface boundary condition indicate a significantly larger portion of the transverse shear load is carried in the bottom layer for this loading condition. For large interface compliance ($\eta = 10^{-7}$), over 250% larger peak interlaminar shear stress is indicated in the bottom layer than the top layer. The bending stress distributions of Figure 4b indicate similar results with about 60% larger peak bending stress occurring in the bottom layer. This difference between the two loading scenarios is significant as typical design and analysis approaches for most aerospace applications is based upon margins substantially less than this nonsymmetric boundary condition effect.

Observation of the interlaminar shear distributions shown in Figure 6a as a function of strip axial coordinate indicates the existence of large stress concentrations near the lower surface reactions and the applied upper center load with minor stress concentrations due to the edge effect in the vicinity of the lower surface. The transverse normal stress distributions shown in Figure 6b indicate similar stress concentrations with magnitudes on the same order as peak bending stresses indicated in Figure 7. This indicates a significant state of biaxial stress exists around the applied load and reactions which, if neglected, could lead to unconservative designs in thick laminates.

The existence of singular points as described by Lu and Liu [5] in which the inplane normal stress and interlaminar shear stress remain unchanged in the presence of interlayer shear slip is evident for the symmetric boundary condition case considered here. The locations for the singular points are near $z = h/6$ and $z = 5h/6$ for inplane normal stress and near $z = h/3$ and $z = 2h/3$ for interlaminar shear stress. These singular point locations were reported by Lu and Liu [5] as well for simple supports with transverse displacement boundary conditions. For the nonsymmetric boundary condition, we realize only one singular point for interlaminar shear stress near $z = 3h/8$ and two singular points for inplane normal stress in the same vicinity as the symmetric boundary condition case.



(a) (b)

Figure 6: Stress Distributions Along the Length of the Nonsymmetrically Constrained Three Layer Strip, Compliant Bond Interface with $\eta = 10^{-8}$: (a) Transverse Shear Stress, (b) Transverse Normal Stress

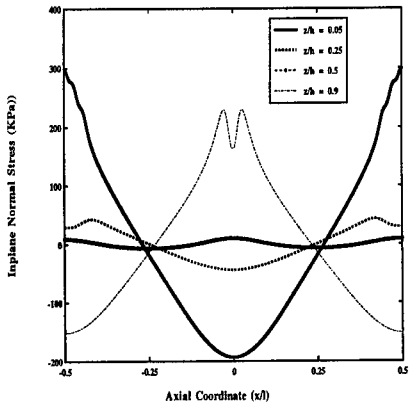


Figure 7: Inplane Normal Stress Distributions Along the Length of the Nonsymmetrically Constrained Three Layer Strip, Compliant Bond Interface with $\eta = 10^{-8}$

6.2 Arbitrary Five Layer Strip: Uniform Load

To investigate the effects of lateral-surface constraint on the stress distribution in an arbitrary laminated strip, here we consider a problem representative of materials utilized in typical aerospace microelectronics applications. The lamination involves a silicon die material on a circuit board bonded to an aluminum heat sink, each joined with compliant bondlines. A range of joint compliance is utilized to assess the trends and impacts of interlayer shear slip on load distribution corresponding to the use of very compliant "thermal" joints. We seek to determine the deformations and stress distribution due to a uniform load of the form:

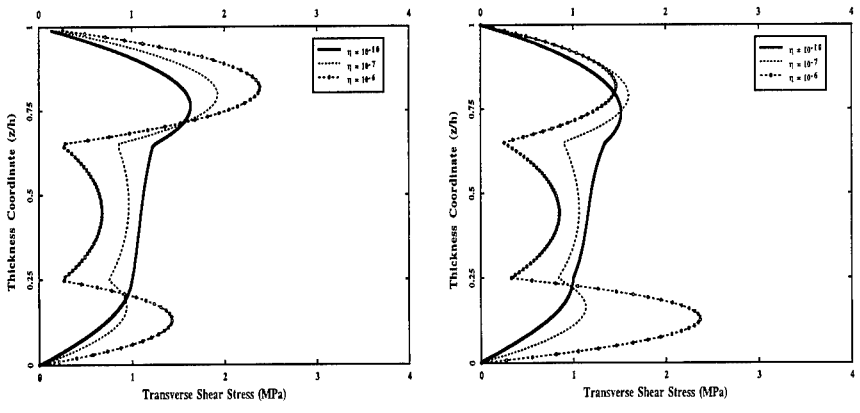
$$P(x) = P_{max} \tag{36}$$

where P_{max} corresponds to a 0.69 MPa (100 psi) pressure load applied over $(|x| \leq \xi_2$. Reaction load is imposed as local transverse pressure at the ends of the strip on upper or lower surfaces of the form:

$$R(x) = \begin{cases} 0 & (|x| < \xi_1) \\ \pm P_{max} \left(\frac{\xi_2}{\xi_2 - \xi_1}\right) & (\xi_1 \leq |x| \leq \xi_2) \end{cases} \tag{37}$$

where $\xi_1 = 0.90\frac{l}{2}$ and $\xi_2 = 0.996\frac{l}{2}$. The lamination is 32.5 mm (1.28 inch) in length with 0.0254 mm (0.001 inch) compliant bondlines and the following layer material properties:

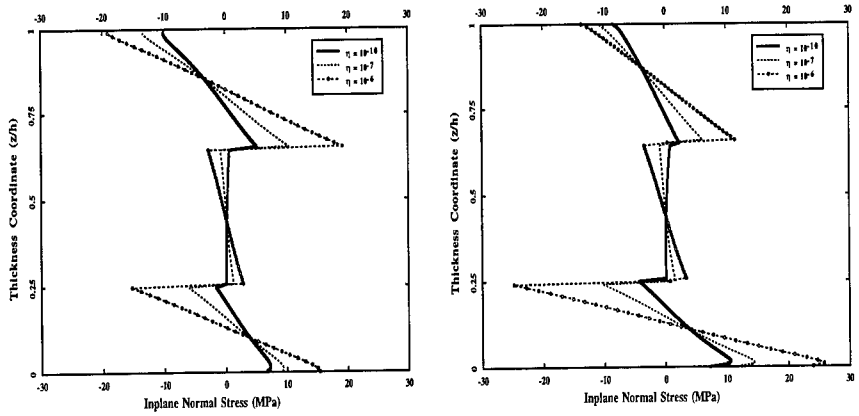
$$\begin{array}{llll} t_1 = 2.032\ mm & E_1 = 68.95\ GPa & \nu_1 = 0.33 \\ t_3 = 3.251\ mm & E_3 = 17.23\ GPa & \nu_3 = 0.45 \\ t_5 = 2.844\ mm & E_5 = 120.65\ GPa & \nu_5 = 0.28 \end{array} \tag{38}$$



(a) (b)

Figure 8: Interlaminar Shear Stress Distributions Through the Thickness of the Arbitrary Strip at $x = -\frac{3}{8}\,l$, S=4: (a) Symmetric End Condition, (b) Nonsymmetric End Condition

The five layer strip described was analyzed by the analytical method described earlier in this paper. The nonsymmetric end condition with transverse reactions taken out on the lower lateral surface is compared to a strictly symmetric end condition having equivalent

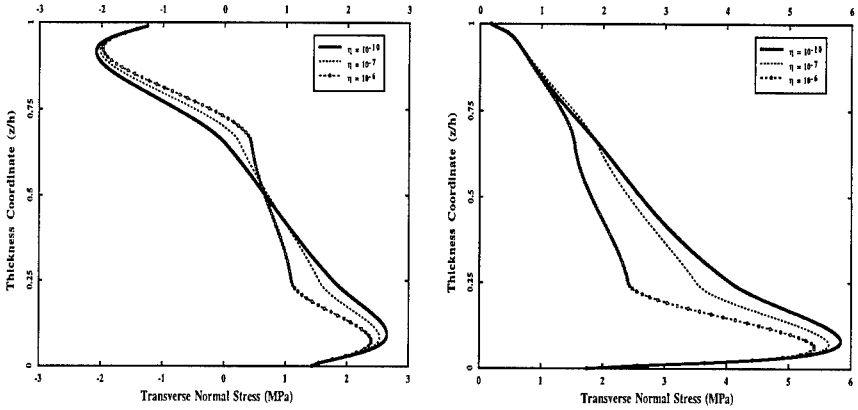


(a) (b)

Figure 9: Inplane Normal Stress Distributions Through the Thickness of the Arbitrary Strip at $x = -\frac{1}{2}\,l$, S=4: (a) Symmetric End Condition, (b) Nonsymmetric End Condition

reactions taken out at both the upper and lower surfaces. A range of interface bond compliance is considered to evaluate the effects of interlayer shear slip on load redistribution. The results are shown in the form of a series of plots for stress distribution as a function of thickness coordinate (z/h). The interlaminar shear and transverse normal stress results of the present solution are shown in Figure 8 and Figure 10 for a range of interface bond compliance of $(10^{-10} \leq \eta \leq 10^{-6})$. The inplane normal stress results of the present solution are shown in Figure 9 for the same range of interface bond compliances. The combination of uniformly applied load and multilayer geometry promoted slow convergence of the theoretical solution with a finite series approximation of $N = 100$ terms required for convergence in a *Mathematica* notebook numerical calculation.

Observation of the results for interlaminar shear distributions in Figure 8a indicates that for symmetric boundary conditions, the shear distribution becomes concentrated in the higher stiffness silicon die as interlayer shear slip develops. The bending stress distributions of Figure 9a for symmetric boundary conditions indicate a significantly larger portion of the bending load taken by the silicon die as well. For large interface compliance $(\eta = 10^{-6})$, about 90% larger peak bending stress is indicated in the silicon due to interlayer shear slip. Comparison with the nonsymmetric end condition results for interlaminar shear distributions in Figure 8b indicates that, similar to the previous symmetric laminate, negligible difference in load distribution exists between the symmetric and nonsymmetric boundary conditions for rigid interfaces. Also, as observed with the symmetric laminate, a marked difference in the load distribution is observed between the two boundary conditions as shear slip develops at the interface. As shear slip develops at the interfaces, the results indicate the load is redistributed to the alumninum heatsink due to the lower surface reaction. For large interface compliance $(\eta = 10^{-6})$, over 60% larger peak interlaminar shear stress is indicated in the heatsink than the silicon die. The bending stress distributions of Figure 9b indicate similar results with more than twice the bending stress occurring in the heatsink than the silicon die. This indicates that a nonsymmetric type of attachment



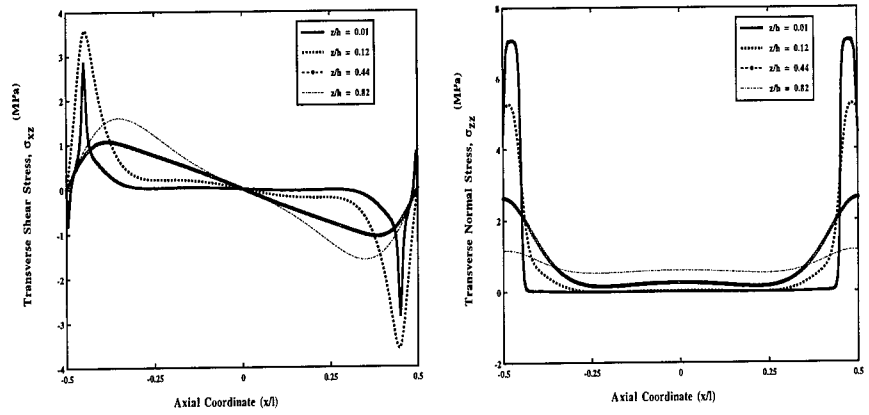
(a) (b)

Figure 10: Transverse Normal Stress Distributions Through the Thickness of the Arbitrary Strip at $x = -\frac{1}{2}\,l$, S=4: (a) Symmetric End Condition, (b) Nonsymmetric End Condition

in this application would actually tend to reduce the stress in the silicon die significantly, particularly once interlayer shear slip develops. By necessity, these types of microelectronics applications typically utilize a nonsymmetric constraint on the heatsink. In addition to the packaging benefits of such a configuration, it appears to have intrinsic benefits relative to maintaining die integrity as well.

Observation of the interlaminar shear distributions shown in Figure 11a as a function of strip axial coordinate indicates the existence of large stress concentrations near the lower surface reactions with minor stress concentrations due to the edge effect in the vicinity of the lower surface. The results of Figure 11b indicate a similar result as with the symmetric lamination example with a significant stress concentration effect extending through half the heat sink thickness. These local stress concentrations are about 50% larger than the magnitudes indicated in Figure 8b away from the vicinity of the reactions. The transverse normal stress distributions shown in Figure 11b indicate similar stress concentrations with magnitudes on the same order as peak bending stresses indicated in Figure 12. This indicates a significant state of biaxial stress exists around the reactions that could have an impact on design of these types of laminates.

The existence of singular points as observed in the previous three layer symmetric example is evident for the symmetric boundary condition case considered here. The locations for the singular points are near $z = h/10$ and $z = 7h/8$ for inplane normal stress and near $z = h/5$ and $z = 7h/10$ for interlaminar shear stress. For the nonsymmetric boundary condition, we realize only one singular point for interlaminar shear stress near $z = h/5$ and two singular points for inplane normal stress in the same vicinity as the symmetric boundary condition case.



(a) (b)

Figure 11: Stress Distributions Along the Length of the Nonsymmetrically Constrained Arbitrary Strip, Compliant Bond Interface with $\eta = 10^{-7}$: (a) Transverse Shear Stress, (b) Transverse Normal Stress

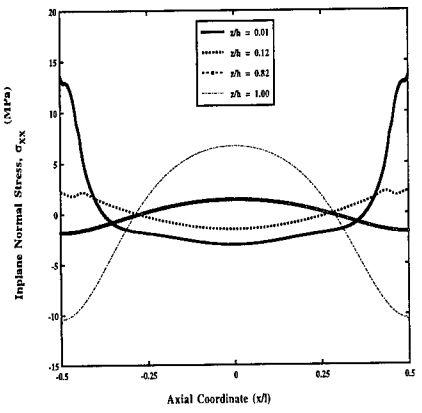


Figure 12: Inplane Normal Stress Distributions Along the Length of the Nonsymmetrically Constrained Arbitrary Strip, Compliant Bond Interface with $\eta = 10^{-7}$

7 Conclusion

In conclusion, a solution has been developed for the plane elastic problem of an isotropic multilayered strip subjected to arbitrarily axial symmetric transverse loads with shear free ends. The solution does not exactly satisfy the free edge normal traction requirement since only resultant force is enforced to zero; however, the solution converges to the particular boundary conditions corresponding to those of enforced zero transverse edge slope. Consequently, the solution is exact for shear free guided ends. The resulting interlaminar stress distributions indicate the effects of nonsymmetric end condition can significantly affect load distribution between the constituent laminae when interlayer shear slip develops due to compliance in the bond joint or loss of bond integrity. These interlayer shear slip effects also may affect the load capacity of thick laminates to a much greater extent than indicated previously in the literature. In fact, the results indicate the impacts of nonsymmetric end constraint in shear slip susceptible designs can be the dominant mechanism limiting the load capability of the structure. However, the solution indicates that these effects can be utilized advantageously as design variables to actually protect fatigue sensitive layers from significant transverse load effects.

References

[1] Bogy, D., 1970, "On the Problem of Edge-Bonded Elastic Quarter-Planes Loaded at the Boundary," *International Journal of Solids and Structures*, Vol. 6, pp. 1287-1313.

[2] Goodman, J.R.,and Popkov, E.P., 1968, "Layered Beam Systems with Interlayer Slip," *Journal of Structural Division*, ASCE, Vol. 24, No. ST11, pp. 2535-2547.

[3] Goodman, J.R., 1969, "Layered Wood Systems with Interlayer Slip," *Wood Science*, Vol. 1, No. 3, pp. 148-158.

[4] Hess,M.S., 1969, "The End Problem for a Laminated Elastic Strip- I. The General Solution," *Journal of Composite Materials*, Vol. 3, pp. 262-280.

[5] Lu,X., and Liu, D., 1992, "Interlayer Shear Slip Theory for Cross-Ply Laminates with Nonrigid Interfaces," *AIAA Journal*, Vol. 30, No. 4, pp. 1063-1073.

[6] Pagano, N.J., 1969, "Exact Solutions for Composite Laminates in Cylindrical Bending," *Journal of Composite Materials*, Vol. 3, pp. 398-411.

[7] Pagano, N.J., 1978, "Stress Fields in Composite Laminates," *International Journal of Solids Structures*, Vol. 14, pp. 385-400.

[8] Reissner, E., 1950, "On a Variational Theorem in Elasticity," *J. of Math. Phys.*, Vol. 29, p. 90.

[9] Reissner, E., 1984, "On a Certain Mixed Variational Principle and a Proposed Application," *International Journal for Numerical Methods in Engineering*, Vol. 20, pp. 1366-1368.

[10] Swett, D.W., and Shiflett, G.R., 1995, "Edge Stresses in Laminated Elastic Strips Subjected to Axially Symmetric Temperature Gradients, Part I: Development of the Theoretical Solution," accepted by ASME *Journal of Electronic Packaging*.

[11] Swett, D.W., and Shiflett, G.R., 1995, "Edge Stresses in Laminated Elastic Strips Subjected to Axially Symmetric Temperature Gradients, Part II: Evaluation of Numerical Results," accepted by ASME *Journal of Electronic Packaging*.

[12] Thompson, E.G., Goodman, J.R., and Vanderbilt, M.D., 1975, "Finite Element Analysis of Layered Wood Systems," *Journal of Structural Division*, ASCE, Vol. 101, No. ST12, pp. 2659-2672.

[13] Toledano, A., and Murakami, H., 1988, "Shear-deformable Two-Layer Plate Theory with Interlayer Slip," *Journal of Engineering Mechanics Division*, ASCE, Vol. 114, No. 3, pp. 605-623.

[14] Vanderbilt, M.D., Goodman, J.R., and Criswell, W.E., 1974, "Service and Overload behavior of Wood Joist Floor Systems," *Journal of Structural Division*, ASCE, Vol. 100, No. ST1, pp. 11-29.

[15] Whitney, J.M., and Pagano, N.J., 1970, "Shear Deformation in Hetergenous anisotropic Plates," *Journal of Applied Mechanics*, Vol. 37, p. 1031.

[16] Whitney, J.M., 1972, "Stress Analysis of Thick Laminated Composite and Sandwich Plates," *Journal of Composite Materials*, Vol. 6, pp. 426-440.

[17] Whitney, J.M., and Sun, C.T., 1973, "A Higher Order Theory for Extensional Motion of Laminated Composites," *Journal of Sound and Vibration*, Vol. 30, p. 85.

[18] Yang, P.C., Norris, C.H., and Stavsky, Y., 1966, "Elastic Wave Propagation in Heterogenous Anisotropic Plates," *International Journal of Solids Structures*, Vol. 2, p. 665.

[19] Yu, Y.Y., 1959, "A New Theory of Elastic Sandwich Plate- One Dimensional Case," *Journal of Appled Mechanics*, Vol. 26, No. 3, pp. 415-421.

EFFECT OF STEADY-STATE OPERATING TEMPERATURE ON POWER CYCLING DURABILITY OF ELECTRONIC ASSEMBLIES

James M. Kallis and Michael D. Norris
Hughes Aircraft Company
P. O. Box 902
El Segundo, California 90245

Abstract: Temperature (thermal) cycling is a major cause of failures of electronic systems. This paper discusses the effect of the steady-state operating temperature, as controlled by parameters such as the coolant inlet temperature, on the fatigue life of electronic assemblies subjected to thermal cycling resulting from off-on power cycling. The power cycling fatigue life depends on the steady-state operating temperature T_s, as well as on the temperature range ΔT in the thermal cycle. The power cycling durability of copper plated through holes in circuit boards is very sensitive to the steady-state PTH operating temperature, because the strain ranges are in the portion of the copper S-N_f (strain vs. number of cycles to failure) curve that is nearly horizontal. If the power is turned on at the same time the coolant flow is started, reducing the steady-state operating temperature can improve the durability. If the coolant is turned on for a substantial period before the power is turned on, reducing the coolant inlet temperature may degrade the durability for operation on hot days.

Key Words: Durability; electronic assemblies; power cycling; steady-state temperature

INTRODUCTION: Temperature is a key contributor to failures of electronic equipment [1, Chapter 4]. Failures result from:

- steady high temperature (hotter than normal room ambient temperature)
- steady low temperature (colder than normal room ambient temperature)
- temperature (thermal) cycling above and below ambient temperature.

Usually electronic equipment operates at temperatures above normal room ambient (20 to 25°C), and the focus is to keep the parts cool. The steady-state operating temperature T_s is controlled by a number of design parameters including, for actively cooled equipment, the coolant inlet temperature [1, Chapter 8].

However, the effect on reliability of the steady-state operating temperature is controversial (see Morris and Reilly [2] and the references cited therein). Some state that:

- microelectronics reliability is independent of, or only weakly dependent on, the steady-state operating temperature T_s
- the dominant temperature dependence is on the temperature range ΔT in a thermal cycle, which affects the number of thermal cycles to failure.

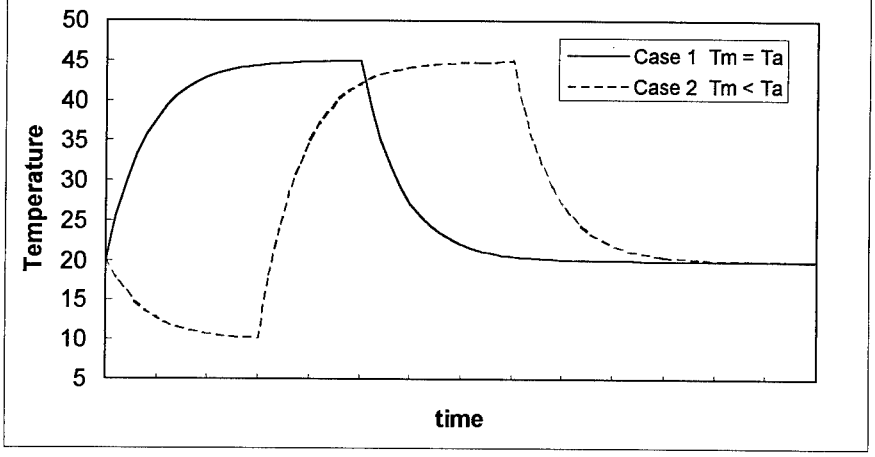
© Copyright 1996
Hughes Electronics

This paper shows that the fatigue life of electronic assemblies, subjected to thermal cycling resulting from off-on power cycling, depends on the steady-state operating temperature T_s, as well as on the temperature range ΔT. The physical basis of the effect of the steady-state operating temperature on the power cycling durability of an assembly is described. The effect is illustrated by a thermal/structural/durability analysis of a representative electronic assembly.

PHYSICAL BASIS: During its service life, the system is powered repeatedly after being "soaked" at the ambient temperature, T_a, and heats up to the steady-state operating temperature, T_s. Each off-on power cycle produces a thermal cycle which damages the equipment. The thermal cycling fatigue damage accumulates and eventually will cause fracture, which in turn will result in an open circuit.

The steady-state operating temperature, T_s, affects the temperature range in each off-on power cycle. The strain range of the materials in the assembly, and thus the damage, increases with increasing temperature range. The number of cycles to failure, N_f, decreases with increasing strain range per cycle.

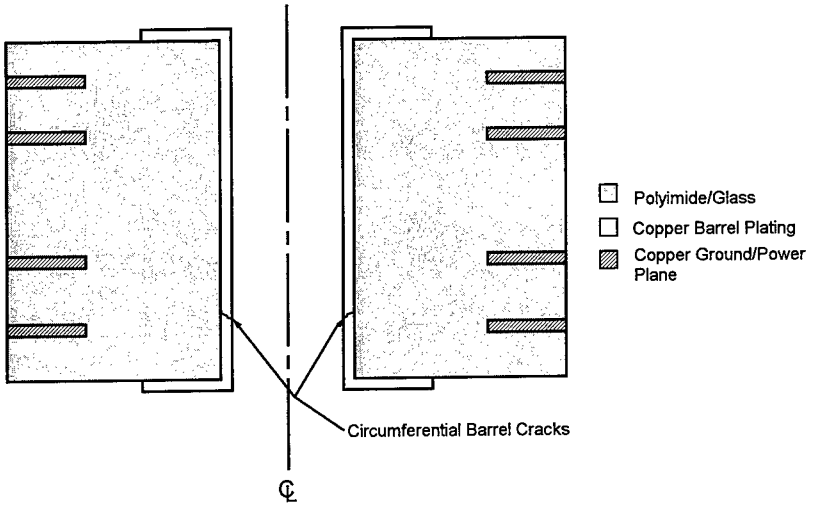
Figure 1. Equipment Temperature vs. Time for Different Coolant Flow Start Times



For actively cooled systems, 1) the power may be turned on at the same time the coolant flow is started or 2) the coolant may flow for a significant period before the power is turned on. In the first case, the equipment may heat up immediately when powered; in that case, the minimum temperature in the power cycle, T_m, is the ambient temperature, T_a. In the second case, the equipment temperature may approach the coolant inlet temperature prior to powering; on a hot day, T_m may be lower than T_a. See Fig. 1.

EXAMPLE - PLATED THROUGH HOLE: Plated through holes (PTHs) electrically connect different layers of a printed wiring board (PWB). An axisymmetric representation of a PTH is shown in Fig. 2.

Figure 2. Cross-Section of a Plated Through Hole



PTH failures during thermal cycling are caused by mismatches in the coefficient of thermal expansion (CTE) between the copper in the PTH barrel and the PWB laminate materials. The PWB laminate materials (polyimide in this case) expand much more rapidly with increasing temperature than the PTH barrel copper. (See Fig. 3.) Typically, the temperature excursions are not large enough to cause failure in a single cycle. Instead, fatigue can cause crack initiation leading to one of the principal electrical failure modes of a PTH - an open circuit resulting from cracking around the circumference of the PTH barrel. These cracks are shown in Fig. 2.

Analytical Approach: This section gives a brief overview of the approach used to evaluate the durability of a PTH.

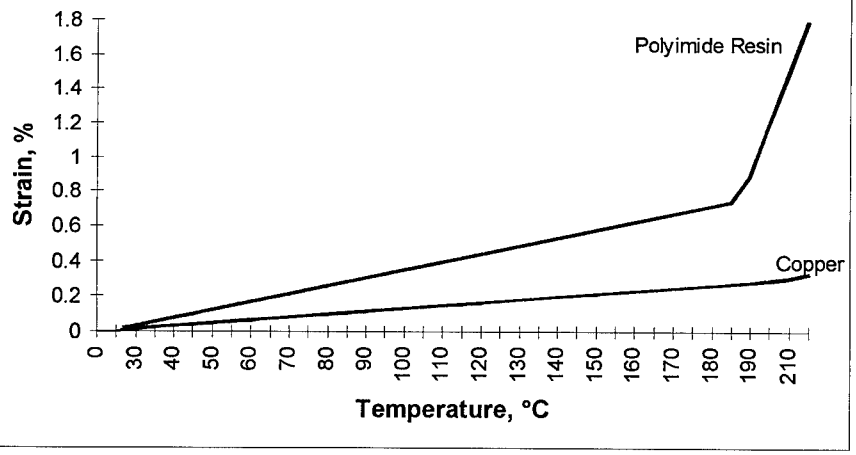
1. *Thermal analysis*: Perform detailed finite-difference analysis [1, Chapter 8] to predict the steady-state operating temperatures within the assembly as a function of coolant inlet temperature. The analysis predicted that:
   - the difference between the steady-state operating temperature of the PTHs (the peak temperature of the PWB surface), T_s, and the coolant inlet temperature is 30°C
   - this difference is independent of the coolant inlet temperature.

2. *Thermomechanical stress/strain analysis*: Perform detailed thermomechanical finite element analysis (FEA) to predict the stresses and strains (elastic + plastic) in the PTH barrel copper as a function of temperature. All finite element models (FEM) are axisymmetric. Copper plasticity is represented by a bilinear kinematic-hardening stress-strain curve. The finite element model (FEM) accounts for the temperature dependence of the properties of copper and the PWB composite. The decrease of copper yield stress with increasing temperature is accounted for. Non-linear strains are calculated at each temperature in the thermal environment of the PTH.

3. *Fatigue analysis*: Predict the thermal fatigue life for the PTH copper using using Engelmaier's modified Coffin-Manson approach, with the maximum equivalent non-linear strains calculated in step 2 as an input. Using Engelmaier's equation [3] (a modified Coffin-Manson approach), the number of cycles to failure expected for each exposure level is estimated. Then the fatigue life resulting from the combinations of individual thermal exposures within the environment is determined using the Palmgren-Miner linear damage accumulation methodology.

Failure modes other than fatigue in the barrel, such as delamination of the PWB or debonding of the barrel and pad from the laminate, are not considered.

Figure 3. Thermal Growth of Polyimide Resin and Electrodeposited Copper



Evaluation of Effect of Copper Plating Quality: Copper plating is assumed to be of high quality without defects that cause localized strain concentrations such as flaws, voids,

localized thickness variations, and lack of adhesion. Variations in plating quality are addressed by use of a quality factor, K_q, discussed below.

Copper quality is represented by K_q and attempts to account for defects in the plating such as flaws, voids, localized thickness variations, and lack of adhesion which induce strain concentrations. The factor $10/K_q$ functions as a strain multiplier:

$$\varepsilon_{effective} / \varepsilon = 10 / K_q$$

Although K_q can range from 0 (infinitely poor plating quality) to 10 (perfect plating quality), it usually falls in the range of 5 to 10. For comparison, a value of $K_q = 7$ results in 43% higher strains than $K_q = 10$.

Fatigue Life Calculation: The maximum equivalent strains calculated in the FEA are used in Engelmaier's modified Coffin-Manson approach to predict fatigue life. Life for each thermal exposure is calculated using these strains in the equation below and solving iteratively.

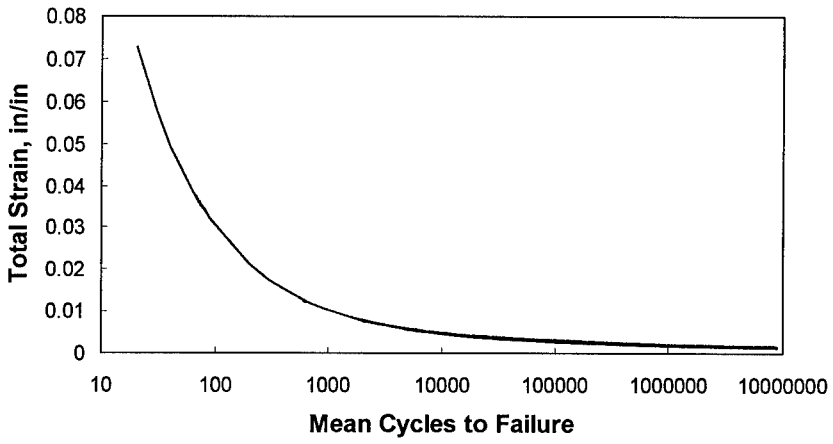
$$N_f^{-0.6} D_f^{0.75} + (S_u / E_{cu}) [\exp (D_f) / 0.36]^{\chi} - \Delta\varepsilon = 0 \qquad (1)$$

where

N_f = expected mean fatigue life N(50%), cycles to failure
D_f = fracture strain, fatigue ductility of barrel copper (baseline = 30%)
S_u = ultimate tensile strength of barrel copper (baseline = 40,000 psi)
E_{cu} = copper modulus of elasticity
$\Delta\varepsilon$ = total cyclic strain range = $| \varepsilon(T_1) - \varepsilon (T_2) |$
T_1 = first temperature of thermal exposure
T_2 = second temperature of thermal exposure
$\chi = 0.1785 \log (10^5 / N_f)$

The relationship of fatigue life and total cyclic strain range in Eq. 1 may be graphically displayed as shown in Fig. 4 for the baseline copper material properties.

Figure 4. Strain Range vs. Mean Cycles to Failure for Baseline Copper Property Values



Evaluation of the Effects of Multiple Exposures: Linear cumulative damage theories can be used to approximate the influence of multiple thermal exposure types. Eq. 2 is the Palmgren-Miner linear damage accumulation expression:

$$D = \sum_{i=1}^{j} \frac{n_i}{N_i} \qquad (2)$$

where

i = exposure type number
j = total number of exposure types
D = damage value
n_i = cycles occurring in exposure i
N_i = life predicted for an exposure i

The Palmgren-Miner theory is an accepted means of combining the effects of multiple fatigue exposures. The relation linearly sums the ratios of experienced cycles to predicted cycles to failure for each of the exposure types within a thermal environment. The cumulative total damage for the environment is then compared to an allowable damage value, usually 0.6 to 1.0.

Results: Calculations were performed for:

- Minimum PTH temperatures, T_m, between -40°C and +40°C
- Steady-state PTH operating temperatures, T_s, between 45°C and 85°C
- coolant inlet temperatures of 15°C and 35°C, which correspond to T_s = 45°C and 65°C, respectively.

The results are shown in Table I and Figs. 5-8: predicted numbers of cycles to failure, N_f. (Note: For lifetimes and cycle rates of military and commercial systems, only values of N_f smaller than about 10^6 are significant. For example, off-on power cycling every hour for 20 years would produce $2X10^5$ cycles. Some of the values of N_f for this example are >> 10^6; however, they illustrate the trends.) Figs. 5 and 6 show N_f vs. T_s for T_m = -40 and 25°C, respectively. Fig. 7 shows N_f vs. T_m for T_s = 45°C and 65°C. For the case in which the power is turned on at the same time the coolant flow is started and the system heats up immediately, $T_m = T_a$. Fig. 8 is the same as Fig. 7 for the limiting case in which the coolant is turned on for a substantial period, and the equipment temperature reaches the coolant inlet temperature, before the power is turned on.

Table I. Example of Effects of Steady-State Operating Temperature on Thermal Cycling Fatigue of Plated Through Holes in Printed Wiring Boards

Minimum PTH temperature, T_m (°C)	Steady-State PTH Operating Temp., T_s (°C)	Copper strain range per cycle, $\varepsilon_{Ts} - \varepsilon_{Tm}$	Thermal cycles to failure, N_f	Ratio, N_f(Ts=45°C) / N_f(Ts=65°C)
-40	45	0.0076	$3.7X10^3$	1.4
-40	65	0.0084	$2.7X10^3$	
10	45	0.0016	$3.4X10^7$	34.
10	65	0.0024	$1.0X10^6$	
40	45	0.0002	$2.5X10^{15}$	1,300,000.
40	65	0.0011	$1.9X10^9$	

Figure 5 Thermal Cycling Fatigue Life vs. Steady-State PTH Operating Temperature
[PWB Soak Temperature (Minimum PTH Temperature) = -40°C]

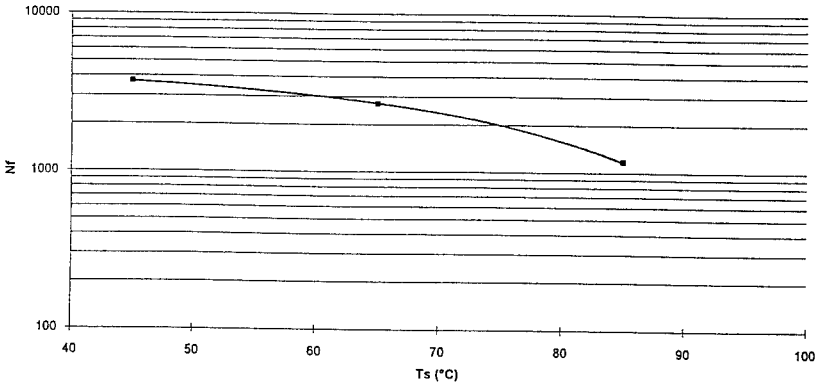


Figure 6 Thermal Cycling Fatigue Life vs. Steady-State PTH Operating Temperature
[PWB Soak Temperature (Minimum PTH Temperature) = 25°C]

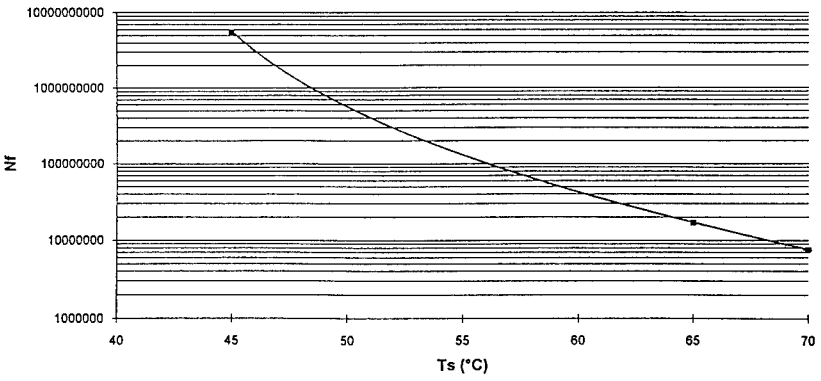


Figure 7 Thermal Cycling Fatigue Life vs. PWB Soak Temperature
(Minimum PTH Temperature)

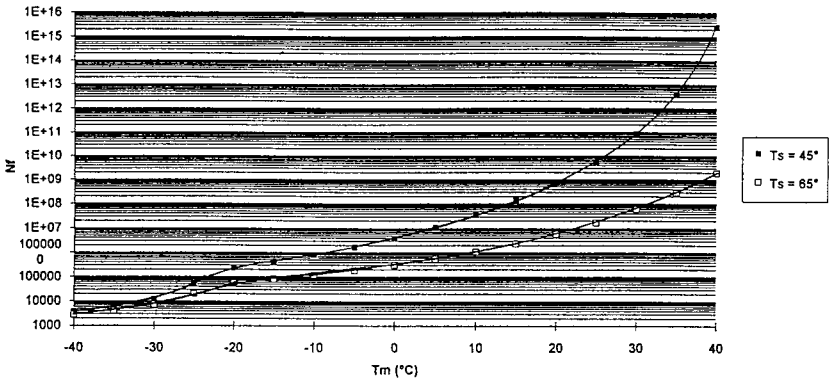
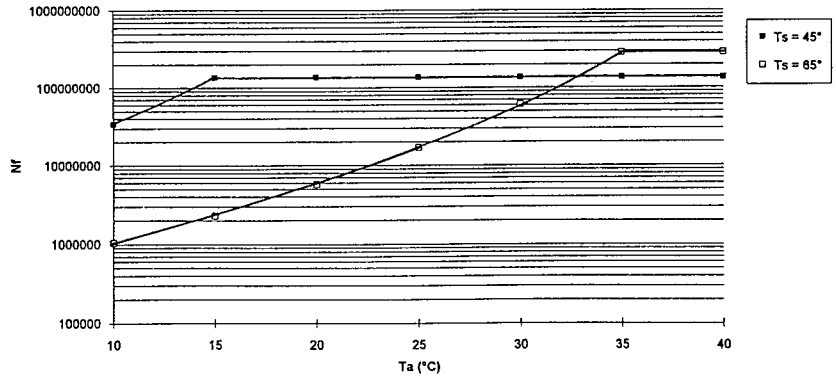


Figure 8 Thermal Cycling Fatigue Life vs. Ambient Temperature for System that Turns on Coolant for Substantial Period before Turning Power On



CONCLUSIONS

- If the power is turned on at the same time the coolant flow is started, reducing the steady-state operating temperature can improve the durability.
- If the coolant is turned on for a substantial period before the power is turned on, reducing the coolant inlet temperature may degrade the durability for operation on hot days. For example, Fig. 8 shows that, for ambient temperatures higher than 33°C, the number of cycles to failure is smaller for a 15°C coolant than for a 35°C coolant.
- The power cycling durability is very sensitive to the steady-state PTH operating temperature, because the strain ranges are in the portion of the copper S-N_f (strain vs. number of cycles to failure) curve that is nearly horizontal. For example, for an initial temperature of -20°C, a 20°C increase in the operating temperature (from 45°C to 65°C) increases the strain by a a factor of 1.3 (from 0.0030 to 0.0038), which decreases the number of cycles to failure by a factor of 4 (from $2.3X10^5$ to $5.6X10^4$).

ACKNOWLEDGEMENTS: Mark Hontz performed the thermal analysis for the example. David Sandkulla introduced to Hughes the approach to the thermomechanical stress/strain analysis and fatigue analysis used in the example.

REFERENCES

1. Morrison, G.N., Kallis, J.M., Strattan, L.A., Jones, I.R., Lena, A.L., MacDiarmid, P.R. and Unkle, C.R., 1995, RAC Thermal Management Guidebook, 1995, Reliability Analysis Center, 201 Mill Street, Rome, NY 13440.

2. Morris, S.F. and Reilly, J.F., MIL-HDBK-217 - A Favorite Target, 1993 Proc. Annual Reliability and Maintainability Symp., pp. 503-509.

3. Engelmaier, W., et al., "Round Robin Reliability Evaluation of Small Diameter Plated Through Holes In Printed Wiring Boards," IPC-TR-579, Institute for Interconnecting and Packaging Electronic Circuits (formerly Institute of Printed Circuits), September 1988.

MFPT

SENSORS TECHNOLOGY

CoChairs: Henry R. Hegner
ManTech Advanced Technology Systems

Christopher P. Nemarich
Naval Surface Warfare Center/Carderock Division

GRATING-BASED OPTICAL FIBER SENSORS FOR STRUCTURAL ANALYSIS

Vikram Bhatia, Vivek Arya, Tiffanie D'Alberto, Daniel Sherr, David Campbell, Marten J. de Vries, and Richard. O. Claus
Fiber & Electro-Optics Research Center, Bradley Department of Electrical Engineering
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0356

Christopher P. Nemarich
Naval Surface Warfare Center, Carderock Division, Code 8530
3A Leggett Circle, Annapolis, MD 21402-5067

Abstract: Long-period gratings are versatile optical fiber sensing elements that are simple and economical to fabricate and demodulate. Long-period gratings are periodic photoinduced structures in fiber cores that couple light from guided to cladding modes. This paper discusses the applications of these devices to strain measurements in high-performance materials and structures. Experimental results from a preliminary loading test carried out on rebar commonly used to reinforce civil structures are presented. The cross-sensitivity of long-period grating-based strain sensors to temperature is analyzed and two methods to overcome this limitation are presented. Furthermore, the application of such gratings to temperature measurements in the presence of actively varying axial strain is discussed. Preliminary results indicate that long-period gratings hold tremendous potential for health monitoring of advanced materials and structures.

Key Words: Long-period gratings; optical fiber sensors; smart materials; smart structures; strain-insensitive gratings; temperature-insensitive gratings

INTRODUCTION: In the past few years fiber optic techniques have emerged as viable alternatives to conventional sensors used to monitor structural health [1,2]. Fiber optic sensors are small in size, simple to multiplex and immune to electromagnetic interference. One common optical fiber sensor used extensively in different applications is the fiber Bragg grating [3]. Bragg gratings are typically formed by side-exposing a germanosilicate fiber to a spatially periodic intensity pattern of ultra-violet (UV) light [4]. The photosensitivity phenomenon modulates the refractive index of the fiber core and causes light at a particular wavelength to reflect back towards the optical source. This wavelength-selective nature of Bragg gratings is a function of the fiber parameters and the spatial period of the index of refraction modulation [3]. An external perturbation, such as axial strain, produces a change in the fiber properties and the grating period, and results in a displacement of the reflected wavelength. For a calibrated sensor, this spectral shift can be employed to measure the magnitude of the applied strain. The advantage of using Bragg gratings is that a number of such devices can be wavelength-division multiplexed on a single fiber to yield a network of distributed sensors [5]. This reduces the cost per channel and results in real-time absolute measurements throughout the lifetime of the host material or structure.

This paper introduces long-period grating sensors as effective tools to monitor the condition of civil structures. The differences between the operating mechanism, performance and fabrication of long-period and fiber Bragg gratings are explained. Long-period gratings are demonstrated as versatile devices that can potentially be used for strain measurement in both surface-mounted and embedded configurations. Experimental results from a laboratory test on a mechanically loaded reinforcing-bar are presented to verify the strain sensing capability of long-period gratings. A major limitation of grating-based sensors is their cross-sensitivity to ambient temperature fluctuations. Two techniques for obtaining temperature-independent strain measurements

using long-period gratings are discussed. The first technique uses specially designed fibers to eliminate the temperature-sensitivity of long-period gratings. The second method uses the multiple bands of a grating to extract information about both strain and temperature changes. This simultaneous measurement scheme is extremely attractive for applications where both thermal and mechanical changes have to be monitored concurrently. A grating with a very small axial strain-induced shift is employed to measure temperature variations in the presence of actively varying strain. Finally, the advantages and limitations of long-period grating-based sensors are discussed.

LONG-PERIOD GRATINGS: The differences between long-period and fiber Bragg gratings can be explained on the basis of the β-plots of Figure 1. Such plots are used to represent the propagation constants (denoted by β) of different modes in an optical fiber [6]. The region with positive propagation constants (β>0) corresponds to modes that propagate in the forward direction (from source to detector) while β<0 includes all reverse- or backward-propagating modes. For power coupling to occur between two modes in a grating with period Λ, the modal propagation constants (denoted by β_1 and β_2) should satisfy the phase-matching condition,

$$\beta_1 - \beta_2 = \Delta\beta = \frac{2\pi}{\Lambda}, \qquad (1)$$

where Δβ is the differential propagation constant.

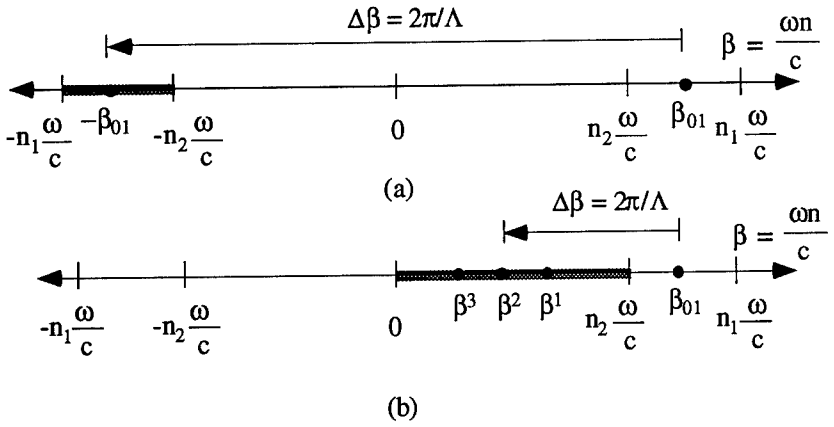


Fig. 1. β-plots for coupling in a (a) fiber Bragg grating and (b) long-period grating [6]. The difference in the grating periods arises from the distinct values of Δβ in the two cases. n_1 and n_2 are the refractive indices of the fiber core and cladding, respectively.

For a fiber Bragg grating (see Figure 1(a)) light is coupled from the forward-propagating fundamental guided mode, with propagation constant $\beta_1=\beta_{01}$, to the reverse-propagating guided mode, with propagation constant $\beta_2=-\beta_{01}$. Such coupling between modes with a relatively large differential propagation constant results in requirements of grating periods of small magnitude (Equation (1)). Typical values of the periodicity of the refractive index modulation for short-period Bragg gratings are about 0.5 μm. Since the propagation constants are strong functions of the operating wavelength, the coupling to the reverse-propagating guided mode is highly wavelength selective. The reflected

wavelength depends on the grating period and the effective index of refraction of the guided mode. Any modulation of the period or effective index shifts the Bragg wavelength to a new value and this concept is commonly employed to implement strain and temperature sensors [3]. Bragg gratings are typically fabricated by impinging the periodic interference pattern of two UV beams, such as that from a phase mask, on the fiber core.

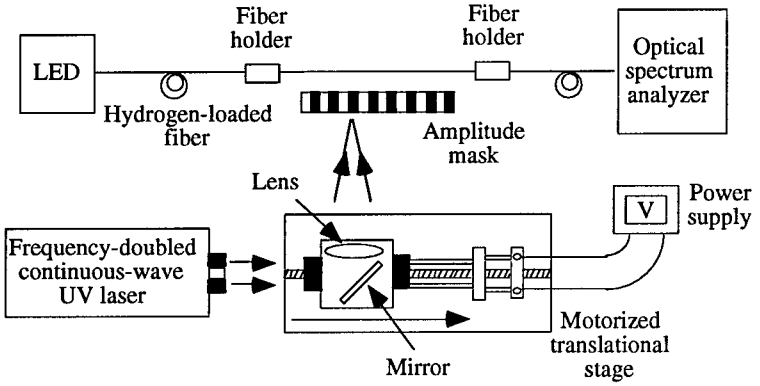


Fig. 2. Set up to fabricate long-period gratings using continuous-wave UV exposure from a frequency-doubled argon-ion laser.

Long-period gratings [6] are also based on the fulfillment of the phase-matching condition between different modes in an optical fiber. As the name suggests, such devices have longer periods than their Bragg grating counterparts, typically of the order of hundreds of micrometers. This results in coupling of light from the fundamental guided mode ($\beta_1=\beta_{01}$) to cladding modes propagating in the same direction, as depicted in Figure 1(b). For an unblazed grating, power is coupled to circularly-symmetric cladding modes of order m ($\beta_2=\beta^m$). Cladding modes result from the trapping of the power radiated by the core at the outer surface of the cladding. Each mode attenuates rapidly due to fiber bends and imperfections and hence produces a characteristic loss band in the grating transmission spectrum. Since the guided mode couples light to a number of cladding modes, the resultant spectrum consists of a series of resonance bands at wavelengths,

$$\lambda^{(m)} = \left|\delta n_{eff}^{(m)}\right| \Lambda , \qquad (2)$$

where $\delta n_{eff}^{(m)}$ denotes the difference between the effective index of the guided mode and that of the cladding mode of order m. External perturbations that change the differential effective index and/or the grating period serve to modulate the location of the resonance bands [7]. The shift in each resonance band due to the same magnitude of a given perturbation is distinct and this concept can hence be used to implement multi-parameter sensors.

Long-period gratings are fabricated by exposing a hydrogen-sensitized germanosilicate fiber to continuous or pulsed UV light through an amplitude mask with rectangular transmittance function of appropriate period [6]. The set up for writing long-period

gratings using continuous-wave irradiation from a frequency-doubled argon-ion laser is shown in Figure 2. The UV beam is scanned along a given length of bare fiber using a lens and mirror combination mounted on a motorized translation stage. The exposure is continued until maximum power transfer occurs from the guided mode to a given cladding mode [6]. The use of different grating periods and beam translation speeds enables one to manipulate the spectral characteristics of long-period gratings. Since amplitude masks are inexpensive, long-period grating fabrication is a relatively economical process. Following UV exposure the gratings are annealed at 150 to 200 °C for 10 to 15 hours to outgas the residual hydrogen and remove the unstable UV-induced defects [8].

STRAIN SENSING USING LONG-PERIOD GRATINGS: This section discusses the strain sensing capability of long-period gratings. A grating was written in Corning SMF-28 fiber with a period Λ=280 μm using the technique described in the previous section. The annealed grating was tested for its response to axial strain using LEDs as optical sources and an optical spectrum analyzer at the output end. Figure 3 shows the shift in two successive attenuation bands of the grating as a function of strain. The unperturbed bands were located at wavelengths 1607.9 nm (A) and 1332.9 nm (B) at room temperature. The wavelength shift coefficients of axial strain for the two bands were measured to be 19.4 nm/%ε and 2.81 nm/%ε, respectively. In comparison, the strain-induced shifts for Bragg gratings varies from 9 to 11 nm/%ε. The large strain coefficients of particular bands of long-period gratings reinforce the potential of using such devices as effective strain measurement components.

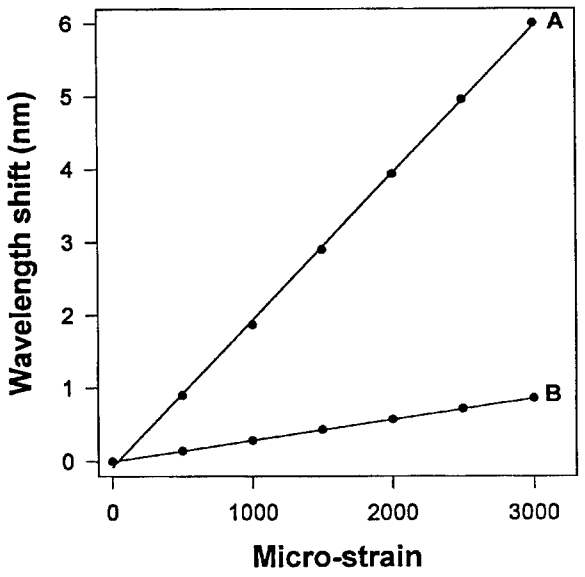


Fig. 3. Strain-induced wavelength shift in two attenuation bands of a long-period grating with Λ=280 μm. The unperturbed location of band A is 1607.8 nm while that of band B is 1332.4 nm.

A long-period grating was attached to the outer surface of a reinforcing-bar (rebar) of 0.5 inch diameter which was then clamped and loaded using an Instron 4468 load frame. An extensometer was collocated on the rebar specimen to obtain an independent strain measurement for the purpose of calibration and comparison. The load on the rebar was increased slowly until the extensometer measured a strain of 1000 με. The shift in the loss band of the grating located at 1636.5 nm was found to be 1.62 nm for 1000 με. Figure 5 (inset) depicts the transmission spectrum of the unperturbed grating and that at 1000 με. The load was then reduced to zero and the rebar sample was loaded in increments corresponding to 200 με on the extensometer digital display. The load on the rebar and the maximum loss wavelength of the grating were recorded in each case. The shift in the resonance band from its unperturbed position was converted into strain using the pre-determined strain coefficient (16.2 nm/%ε). Figure 4 shows the measured strain versus the applied load for this experiment. The rebar was loaded cyclically and consistent measurements were obtained for each run. A fiber Bragg grating adjacent to the long-period grating was similarly calibrated and tested and excellent correlation was obtained between the calculated strain and extensometer data. The strain rate for both the systems is limited by the time taken by the optical spectrum analyzer to complete one scan. This preliminary test reveals the feasibility of using long-period gratings for strain measurement in components and materials used to fabricate large civil structures.

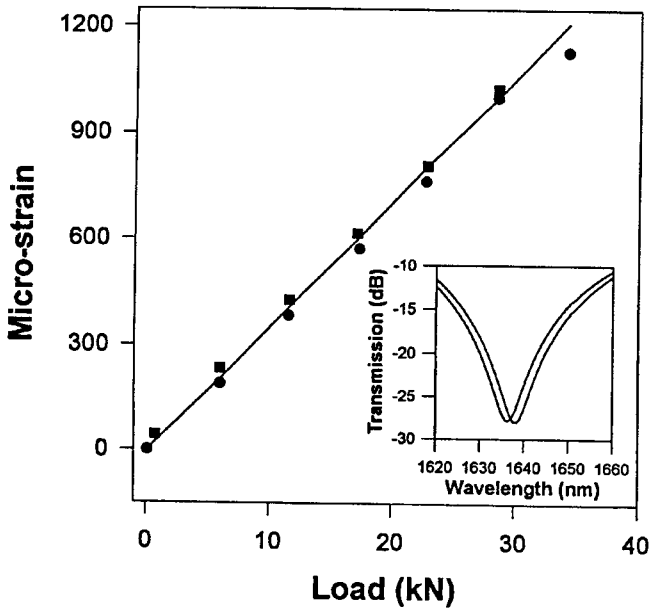


Fig. 4. Measured strain as a function of the load on the reinforcing-bar. The circles are the data from the grating for increasing load while the squares are for decreasing load. The line represents the strain measurement using a co-located extensometer. The inset depicts the transmission spectrum of the grating for an unperturbed case (left) and for 1000 με (right).

TEMPERATURE-INDEPENDENT STRAIN MEASUREMENTS: In the previous section the strain sensing potential of long-period gratings was demonstrated. These experiments were carried out in a laboratory where the room temperature could be controlled to within a few degrees Celsius. In actual applications, the temperature in the sensing region could vary by several tens of degree Celsius. Thus the cross-sensitivity to temperature can severely limit the sensing capacity of long-period gratings [7]. For example, the average temperature-induced wavelength shifts for the two bands of the grating in Figure 3 were found to be 0.093 nm/°C (A) and 0.049 nm/°C (B). These coefficients are much larger than those of standard Bragg gratings (0.011 to 0.013 nm/°C) and hence the thermal cross-sensitivity is a much bigger concern in long-period gratings. In this section two methods to obtain temperature-independent strain measurements are presented.

The first technique employs temperature-insensitive long-period gratings recently proposed by Judkins *et al.* [9]. They divided the temperature-induced shift into a material contribution (change in the differential effective index) and a waveguide contribution (change in grating period) and designed a fiber which counterbalanced these two effects [9]. A grating written in the fiber with special refractive index profile (Λ=166 μm) was tested for its response to temperature and axial strain. An average temperature coefficient of -0.25 nm/100 °C was obtained for this grating. The strain-induced shift for this grating was found to be -27.25 nm/%ε indicating that such devices retain their strain measurement capability despite the reduced temperature sensitivity. To determine the temperature cross-sensitivity at different strain levels, the grating was strained to discrete values and then heated from 25 °C to 125 °C. The temperature coefficients at 500 με and 1000 με are about -0.14 and -0.11 nm/100 °C, respectively. Thus such gratings have almost negligible cross-sensitivity to temperature at different magnitudes of axial strain.

The second method to obtain accurate strain measurements in the presence of ambient temperature fluctuations uses the differential modulation of the attenuation bands of a long-period grating. As shown in Figure 2 the strain response of a band is a function of the order of the corresponding cladding mode. If a grating is perturbed by a strain $\Delta\varepsilon$ and temperature ΔT the wavelength shift in any two resonance bands can be used to separate the two perturbations. For example, if A and B are the temperature and strain coefficients of one band and C and D that of the other band, we obtain,

$$\Delta\lambda_1 = A\,\Delta T + B\,\Delta\varepsilon\,, \tag{3}$$

and,

$$\Delta\lambda_2 = C\,\Delta T + D\,\Delta\varepsilon\,. \tag{4}$$

where, $\Delta\lambda_1$ and $\Delta\lambda_2$ are the cumulative shift due to strain and temperature in the unperturbed bands located at λ_1 and λ_2, respectively. A typical calibration process involves determining the values of the parameters A, B, C and D by independently subjecting the grating to axial strain and temperature. During the actual experiment both strain and temperature are applied simultaneously and the pair of equations (3) and (4) is solved to obtain the unknowns, $\Delta\varepsilon$ and ΔT. It is found that cross-sensitivities and non-linearities make A,B,C and D functions of $\Delta\varepsilon$ and ΔT and hence result in two higher order simultaneous equations. These equations may be solved using an appropriate mathematical software on a personal computer.

To verify the validity of this scheme the grating in SMF-28 fiber (Λ=280 μm) was first calibrated and then heated to discrete temperatures. At these temperatures, the grating was strained and the shift in the two resonance bands was acquired by a personal computer interfaced to the optical spectrum analyzer. The two simultaneous equations

were then solved and values of $\Delta\varepsilon$ and ΔT were obtained. Figure 5 shows the comparison between the actual and calculated values of strain and temperature for this grating. The strain was measured from 0 to 2000 $\mu\varepsilon$ with a maximum error of 38 $\mu\varepsilon$ while the largest error in temperature was 0.83 °C for a range from 23 °C to 55°C. Thus this technique is attractive for applications where both strain and temperature have to be monitored concurrently. The use of a single sensor for strain and temperature measurements reduces the cost of installation and maintenance. The limitation of this scheme is that at present it can only be used for static or quasi-static measurements due to the time taken by the spectrum analyzer to complete two scans per sensing cycle. This drawback may be overcome by using spectrum analyzers with higher scanning rates or by employing modified demodulation schemes.

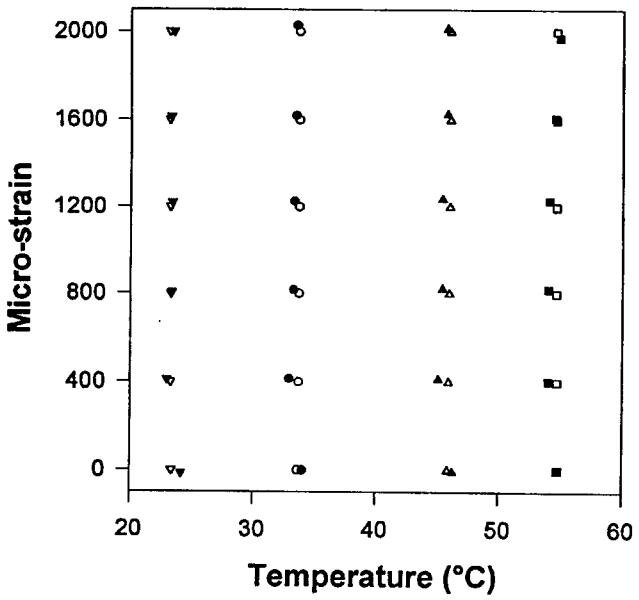


Fig. 5. Actual data (unfilled symbols) and calculated values (filled symbols) for simultaneous strain and temperature measurement using a long-period grating. The grating was heated to discrete temperatures and then axially strained.

STRAIN-INDEPENDENT TEMPERATURE MEASUREMENTS: In some health monitoring applications, it is necessary to measure the temperature in the presence of mechanical changes such as strain. Although both strain and temperature can be determined concurrently using the technique described in the previous section, the development of fiber optic sensors that are immune to axial strain but sensitive to temperature variations can significantly simplify the system operation. Experimental results for measuring temperature using a long-period grating that has a negligibly small strain coefficient are presented.

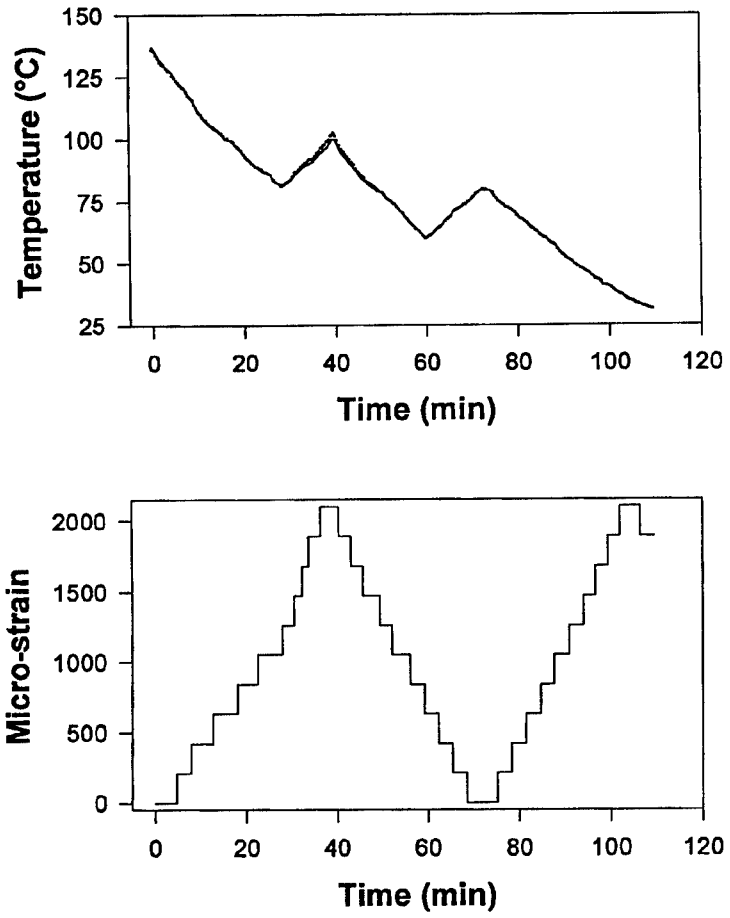


Fig. 6. Temporal variation of temperature (top) and strain (bottom) acting on the grating in Flexcor fiber. The symbols in the top plot represent the temperature measured by employing the calibrated grating.

An overcoupled long-period grating with Λ=340 μm was fabricated in Corning 1060 nm Flexcor fiber and tested for its strain response. The resonance band at 1257.9 nm was found to have a very small strain coefficient of -0.4 nm/%ε at room temperature. A linear curve fit to the temperature-induced wavelength shift yielded a coefficient of 4.33 nm/100°C for this loss band. To test the temperature measurement capability of this grating, it was heated to 140 °C and then cooled slowly, as shown in Figure 6. The axial strain on the grating was varied concurrently between 0 and 2100 με in discrete steps of 210 με. The finite strain coefficient of the grating was neglected for temperature measurement calculations. To test the versatility of the system, the grating was heated

twice during the cooling cycle (Figure 6). The average temperature measurement resolution of the system was found to be 0.8 °C while the maximum error in a calculated value was 2.2 °C over a 110 °C range. The error can be reduced by fabricating gratings with periods that result in bands with identically zero strain-induced shifts. This test suggests that such "strain-insensitive" long-period gratings can be employed to track the ambient temperature in the presence of random axial strain. Additionally, such gratings can be used to extend the temperature dynamic range [10] in the simultaneous measurement system discussed in the previous section.

DISCUSSION: Long-period gratings are versatile components that can be used for single- or multi-parameter measurements. These gratings are simple and economical to fabricate and can be implemented with simple demodulation schemes. For example the LED source at the input can be replaced with a laser diode centered at a wavelength that undergoes an intensity change on being subjected to the desired perturbation [7]. At the output a photodetector can be used to obtain the magnitude of the intensity modulation and hence the value of the applied perturbation. This intensity-based system is expected to be sensitive to bends in the fiber and fluctuations in the source and hence a reference signal might be needed to reduce these undesired effects.

Due to their wavelength selectivity long-period gratings can be multiplexed to produce a series of point sensors on a large structure. The wavelength-division multiplexing of these sensors is limited by the broad spectral width of the resonance bands and possible cross-talk between the multitude of bands of different gratings. The grating bandwidth can be manipulated by varying its length and peak UV refractive index modulation in the core while the wavelength separation between the bands can be controlled by using appropriate grating periods [6]. A number of sensors can also be time-division multiplexed by using the simple demodulation technique described above in conjunction with additional optical components.

The attachment of long-period gratings to structures can be done using techniques similar to those used for fiber Bragg gratings. Although the band-sensitivity of long-period gratings is a critical issue, for most practical applications the sensing region can be expected to be straight during the measurement process. As described earlier, the cross-sensitivity to temperature changes can be overcome using specially designed fiber or by employing the differential shift of two resonance bands of the same grating. Long-period gratings have previously been demonstrated for detecting corrosion in metals [11]. Other applications to "smart" materials and structures are presently under investigation.

CONCLUSIONS: Long-period gratings are promising candidates for providing information about the health of high-performance materials and structures. Preliminary results from strain tests on such devices and their application to loading tests on a reinforcing-bar were presented. Methods to overcome the thermal cross-sensitivity of the strain sensors were discussed. Such techniques would enable the use of long-period gratings in applications where ambient temperature fluctuations significantly influence the employment of other strain sensors. Issues such as sensor demodulation and multiplexing were also addressed briefly. In summary, long-period gratings provide novel methods to measure *in-situ* the condition of civil structures and hold potential as powerful tools for reducing maintenance expenses and for preventing structural failures.

ACKNOWLEDGMENTS: The authors wish to thank Dr. A. M. Vengsarkar at Bell Laboratories, Lucent Technologies for providing the temperature-insensitive long-period gratings.

REFERENCES:

1. E. Udd, *Fiber Optic Smart Structures*, John Wiley and Sons, Inc., New York 1995.
2. J. Dakin and B. Culshaw, *Optical Fiber Sensors: Principles and Components*, Artech House, Boston, 1988.
3. G. Meltz, J. R. Dunphy, W. H. Glenn, J. D. Farina, and F. J. Leonberger, "Fiber optic temperature and strain sensors," *Proc. SPIE Conference on Fiber Optic Sensors II*, vol. 798, 104, 1987.
4. G. Meltz, W. W. Morey and W. H. Glenn, "Formation of Bragg gratings in optical fibers by transverse holographic method," *Optics Letters*, vol. 14, 823, 1989.
5. W. W. Morey, J. R. Dunphy, and G. Meltz, "Multiplexing fiber Bragg grating sensors", *Proc. SPIE on Distributed and Multiplexed Fiber Optic Sensors*, vol. 1586, 216, 1991.
6. A. M. Vengsarkar, P. J. Lemaire, J. B. Judkins, V. Bhatia, J. E. Sipe, and T. E. Ergodan, "Long-period fiber gratings as band-rejection filters," *Journal of Lightwave Technology*, vol. 14, 58, 1996.
7. V. Bhatia and A. M. Vengsarkar, "Optical fiber long-period grating sensors," *Optics Letters*, vol. 21, 692, 1996.
8. P. J. Lemaire and T. Ergodan, "Hydrogen-enhanced UV photosensitivity of optical fibers: mechanisms and reliability", *Proc. OSA Topical Meeting on Photosensitivity and Quadratic Nonlinearity in Glass Waveguides*, paper SuA4, 78, 1995.
9. J. B. Judkins, J. R. Pedrazzani, D. J. DiGiovanni, and A. M. Vengsarkar, "Temperature-insensitive long-period fiber gratings," *Proc. Conference on Optical Fiber Communication*, postdeadline paper PD1, 1996.
10. V. Bhatia, D. Campbell, R. O. Claus, and A. M. Vengsarkar, "Simultaneous strain and temperature measurement using long-period gratings," Submitted to *Optics Letters*, 1996.
11. J. A. Greene, K. A. Murphy, T. A. Tran, V. Bhatia, and R. O. Claus, "Grating-based optical fiber corrosion sensors," *Proc. SPIE Smart Sensing, Processing, and Instrumentation*, vol. 2718, paper A-19, 1996.

HELICOPTER GEARBOX VIBRATION MEASUREMENT WITH THE ROBUST LASER INTERFEROMETER

H. Michael Holland
Theodore C. Goodenow
Robert L. Shipman

Epoch Engineering, Incorporated
806 West Diamond Avenue, Suite 200
Gaithersburg, MD 20878

Abstract: Epoch Engineering, Inc. (EEI) has developed the Robust Laser Interferometer (RLI) which is able to make high quality vibration measurements across a wide bandwidth and with a great dynamic range, continuously. Overall, the RLI is able to measure vibrations of .1Å (peak to peak) in the presence of vibrations of .1m (peak to peak). Bandwidth extends from fractional Hertz to 64 KHz and can be increased to 256 KHz. In this paper, EEI describes its successful effort to measure the vibrations of two helicopter (SH-60 and H-53E) main gearboxes undergoing full power testing on test stands. These measurements are compared to simultaneous accelerometer system measurements. The RLI measured vibrations in a line-of-sight, non-contact fashion with several advantages over the accelerometer system. These include:

(1) Selected measurement locations independent of mechanical attachment points;
(2) Improved low frequency (<50 Hz) performance, including a noise floor up to 50 dB lower than the accelerometer system at low frequencies, revealing low frequency vibrations of known mechanical origin not revealed by the accelerometer system;
(3) Extended high frequency performance, from 8 KHz to 64 KHz, without mounted resonance or other difficulty; and
(4) Demonstrated dynamic range greater than 150 dB (acceleration.)

Key Words: Condition assessment; Helicopter vibration measurement; Low frequency vibration measurement; Machinery diagnostics

Introduction: In this paper, Epoch Engineering Inc. (EEI) describes its successful Phase I Small Business Innovation Research (SBIR) effort to measure the vibrations of two helicopter main gearboxes undergoing full power testing on the test stand at the Naval Aviation Depot, North Island Detachment, Pensacola, FL. These measurements were made using EEI's own, newly developed, Robust Laser Interferometer (RLI) and, in most cases, compares these measurements to accelerometer system measurements taken at the same locations, often simultaneously. In this effort, the RLI demonstrated the ability to measure vibrations in a line-of-sight, non-contact fashion with several distinct advantages over the accelerometer system. These include:

1. At frequencies where the accelerometer system is well suited and

provides reliable measurements, in this case from approximately 50 Hz to 8 KHz, the RLI non-contact readings and the accelerometer readings matched extremely well.

2. At lower frequencies, in this case below approximately 50 Hz and especially below 10 Hz, where the accelerometer system is limited in its ability to provide a true measure of acceleration, the RLI measurements reveal a number of vibrations that correspond to known excitation frequencies originating within the machinery and that were not detected by the accelerometer system. The difference in the noise floor of the accelerometer system and the RLI at frequencies below 10 Hz approaches 50 dB in acceleration.
3. At higher frequencies, in this case from 8 KHz to 64 KHz, the RLI provided measurements that are faithful to the vibration of the structure. There is no indication of mounted resonance or other contact mounting problem.

Discussion: The two helicopter main gearboxes upon which measurements were taken were the SH-60 and the H-53E. The SH-60 was tested during an "A" test after the completion of overhaul and test points included all tested power levels. Measurements were taken on the gearcase near the port and starboard engine inputs and near the center of the gearcase between the two engine inputs. A similar set of measurements at similar test conditions was conducted on the H-53E main gearbox. Additional test points in the vicinity of the accessory gearbox were also taken in the H-53 data set. Data was typically taken in both one second and sixteen second sample times.

EEI found that its RLI system makes non-contact vibration readings of helicopter gearboxes that are comparable to accelerometer readings in the worst case and considerably better than accelerometer readings in some frequency ranges of interest. The ability to measure vibrations of known mechanical origin within the gearboxes in frequency ranges where accelerometer measurements are not reliable has been demonstrated.

In order to understand the objective of this measurement effort it is necessary to understand that a conventional (accelerometer based) vibration measurement system functions within several limitations. Setting aside mounting and attachment issues, the accelerometer based system begins with a mass mounted on a transducing element. Simplistically, the mass tends to remain stationary due to its inertia while the accelerometer casing vibrates around it, and the transducing element produces the signal. This transduction has a measured non-linearity over a specific frequency range. This non-linearity is generally specified as 1% and exists over a range from low frequency, often 10 Hz or lower, to a frequency that approaches some fraction of the mounted resonance of the device. This range is often referred to as the linear range of the accelerometer. This low voltage output, 99% faithful to the input within the specified limits of the device, is then pre-amplified and sent to an Analog-to-Digital Converter (A→D) and to a data logger. Pre-amplifiers generally do not

contribute additional artifacts. The A→D is selected to be large enough to preserve the signal. Often, at this stage, filtering is used in order to limit the logged signal to frequencies where the measurement system is considered "linear."

Reconsidering, however, all artifacts of the 1% non-linearity of the accelerometer are included in the accelerometer signal output. These artifacts may have been generated by effects that are outside the "linear" frequency range of the accelerometer, but when the artifacts themselves occur at frequencies where the accelerometer is considered "linear," the artifacts are carried along as true vibrations.

The investigation attempted to demonstrate that the non-linearity of a quality accelerometer, properly installed and operated, on a machine with a complex and high-amplitude vibration pattern, limited the ability of the system to accurately report vibration levels which were of true mechanical origin within the machine and which could be accurately reported by the Robust Laser Interferometer. The principal differentiating factor is the relatively low (.01% - .001%) non-linearity of the RLI.

Accelerometer specifications:

Specifications	Model	Units
Dynamic		
Sensitivity, ±5%, 25°C	10	mV/g
Acceleration	500	g peak
Amplitude Nonlinearity	1%	
Frequency Response:		
±5%	3.0 - 10,000 Hz	
±10%	2.0 - 12,000 Hz	
±3 dB	1.0 - 15,000 Hz	
Resonance Frequency, mounted, nominal	32 KHz	
Transverse Sensitivity, maximum	5% of axial	
Electrical: Power Requirement:	18 - 30 VDC	
Voltage Source	2 - 10 mA	
Current Regulating Diode		
Electrical Noise, equiv. g, nominal:		µg
Broadband 2.5 Hz to 25 KHz	100	µg/√Hz
Spectral 10 Hz	12	µg/√Hz
100 Hz	3.0	µg/√Hz
1,000 Hz	1.0	µg/√Hz
10,000 Hz	0.6	µg/√Hz

Robust Laser Interferometer Specifications:

Input Channel:

1. Single input channel provides continuous measurement bandwidth from pure displacement (direct current or D.C.) to 128 KHz.
2. Measures either displacement or velocity (selectable). (Velocity is differentiated with respect to time to provide acceleration.)

3. Provides high dynamic range. For velocity or displacement measurement equivalent to 24 bit A to D conversion. Demonstrated continuous dynamic range greater than 150 dB.

Digital Signal Processing:

1. Digital conversion of signal greater than 50 bits.
2. 133 MHz processing clock.
3. Continuous data acquisition up to 128 KHz (dependent upon required data processing.)
4. Time series data sampling rates to 256 KHz (with 1 Hz resolution[1].)
5. Spectral data to 128 KHz (1 Hz resolution) a maximum resolution of 1/32 Hz (to 8 KHz).

Sensor Non-Linearity: The output of a contact sensor mounted on a vibrating object provides an output signal that approximates the input forces. In the sense that the response curve of even the best contact sensor (in this case used normally to mean accelerometer or velocimeter) is not a straight line but typically has some curvature and deviates from a straight line no more (in amplitude) than 1 percent, these sensors are said to have a non-linearity that is less than 1 percent.

The force input to the sensor is itself often non-linear in that the physical item being measured, especially if it is of complex mechanical construction and has many vibration sources at different frequencies, is excited at both the fundamental frequencies of the noise sources but also at frequencies caused by the inter-modulation of these driving frequencies. One well-recognized appearance of these inter-modulation frequencies are "sidebands," which occur typically when a low frequency and higher frequency signal intermodulate and the intermodulations are most easily recognized by their close spacing around the higher frequency signal. Consider a simple case for a non-linear measurement system, such as a forcing function F, where:

$$F = F_1 \sin(\omega_1 t) + F_2 \sin(\omega_2 t)$$

where F_1 and F_2 are the amplitudes of two sinusoidal forces at frequencies ω_1 and ω_2. A linear system would yield amplitude measurements only at f_1 and f_2, that is where $\omega_1=2\pi f_1$ and $\omega_2=2\pi f_2$, but the non-linear system reports amplitudes at the

1. The RLI is currently configured to provide up to 1 Hz resolution for time series data at bandwidth of 256 KHz. This capability can be adjusted to, for example, provide 512 KHz bandwidth at 2 Hz resolution, or 128 KHz bandwidth at 1/2 Hz resolution. The resolution can be as fine as 1/32 Hz, resulting in bandwidth of 8 KHz. For spectral data, the bandwidth is halved, so spectral data to a bandwidth of 128 KHz, with 1 Hz resolution, is possible. Similar proportions for other resolutions as above apply. For example, a 32 second data acquisition time could be used to provide 1/32 Hz resolution for a bandwidth to 4 KHz in order to provide high resolution spectral data.

following frequencies:[2]

$$f_1,\ 2f_1,\ 3f_1,\ f_2,\ 2f_2,\ 3f_2,\ f_1+f_2,\ f_1-f_2,$$
$$2f_1+f_2,\ -2f_1+f_2,\ f_1+2f_2,\ \textit{and}\ f_1-2f_2\ (\textit{see footnote}\ 2).$$

These additional products are intermodulation distortion products, or untrue artifacts of the sensor, and do not represent true vibrations of mechanical origin. There are three aspects of this result to consider:

a. The intermodulation products created by the mechanical system, such as sidebands or other modulations, are true vibrations (at the input to the sensor); and therefore, represent the conditions and effects of the mechanical systems.

b. The non-linearities of the sensor create additional products, called intermodulation distortion products, that add to the true intermodulation products and create the appearance of other intermodulation products that were not present in the mechanical system. These products are artifacts of the sensor system non-linearity.

c. The highlighted difference frequencies mentioned above provide insight into the customary low frequency limitation of accelerometer systems. When measuring a complex mechanical component with many driving frequencies, there are many intermodulation distortion products. The additive products tend to spread out in the spectrum while the difference products tend to collect at the lower frequencies. This collection of the difference products is increased because the negative components are also reported as positive components of the same magnitude (referring to the footnote discussion). The collection of these products creates an apparent noise floor that is higher than the true noise floor of the machinery, especially at the lower frequencies. If a certain low frequency is strongly represented as a true mechanical intermodulation product of higher frequencies, such as the previously mentioned sideband of a running speed around a gear mesh frequency, then the sensor system over-reports the amplitude at the lower frequency (that is, in the example used, between the gear mesh frequency and the sideband) due to the sensor system's creation of the intermodulation distortion products. In this sense, when measuring vibrations of a complex mechanical system with many driving frequencies, there is some point in the spectrum plot below which the amplitudes reported by a system of even 1 percent non-linearity should not be trusted even if their amplitude is above the apparent noise floor. The true driving amplitude at that frequency is likely to be lower and is being over-reported because of the intermodulation distortion products of the sensor system.

---

2. The difference frequencies (those with minus signs) are of special interest, since $\cos(-\omega t)=\cos(\omega t)$ and $\sin(-\omega t)=-\sin(\omega t)$, the term $-\omega t$ can be expressed as $+\omega t$ (i.e. F_1-2F_2 is equivalent to $2F_2-F_1$.) This means that those frequencies created by the difference of two closely spaced driving frequencies have an even greater tendency to collect in the low frequency region.

Additional Factors Affecting Measurements:

Transverse Sensitivity - No measurements were made to attempt to determine the impact of the accelerometer system transverse sensitivity. The accelerometers used have a reported transverse sensitivity of 5 percent. The RLI, when interrogating a surface that is planar from a perpendicular situation, has a very small transverse sensitivity. It is considered likely, but not demonstrated in this effort, that some of the difference in performance at lower frequencies between the accelerometers and the RLI is as a result of transverse sensitivity.

Noise Floor and Signal-to-Noise Ratio (SNR) - The noise floor of any measurement system using an Analog to Digital (A→D) converter is limited by the number of bits in the converter. In this sense, an A→D converter of 12 bits is able to record a dynamic range of 20 log (2^{12-1}) dB = 66 dB (one bit being reserved for the + or - sign.) Similarly, a 16 bit converter can record signals that differ in amplitude by 90 dB. This is one fashion in which the dynamic range of most accelerometer system are limited, but it is not a functional limitation for most accelerometer systems because their linearity (as developed in earlier discussions) is so limiting that additional A→D converter size does not help the overall system performance. In the case of the acceleration measurements made here, the accelerometers were used with a 16 bit A→D converter and crisp digital filtering. In this sense, all attempts were made to make accelerometer systems perform as well as they could.

The RLI, because of the large inherent dynamic range (the RLI has demonstrated the ability to measure vibrations differing in amplitude from .1 meters (four inches) to .00000000001 meters (.1Å) peak-to-peak, simultaneously) is afforded a larger A→D converter. EEI, in discussions with helicopter-experienced instrumentation engineers, had been warned to be prepared for acceleration levels up to 1000 g's. While no levels higher than a few hundred g's were experienced, the system was configured to be able to handle 1000 g's (+ 60 dB) while still providing measurement capability to below .000001 g (- 120 dB). At very low frequencies (< 2 Hz), noise floors as low as .000001 g (-120 dB) were seen on both the SH-60 and the H-53E gearboxes.

Measurement Results (SH-60 Main Gearbox)

The presented measurement set for this condition, included as Figure 1, shows the 16-second sample time measurements of the accelerometer above the simultaneous measurement of the RLI. In the interest of brevity, these plots are only shown to 96 Hz, which also displays the range where the RLI measurements have a substantially lower noise floor than the accelerometer measurements. The first seven harmonics of the main rotor rotational speed (4.29 Hz) are clearly present in the RLI signal. Below, in tabular format are selected peaks that appeared at least 58 percent of the time. The RLI table shows the first (4.29 Hz), second (8.60 Hz), third (12.90 Hz), fourth (17.20 Hz), fifth (21.50 Hz), and sixth (25.79 Hz) harmonics of main rotor rotational vibration, along with other identifiable signals such as the Main Bevel Gear

rotational (20.09 Hz). Referring back to the earlier discussion concerning the reason that a non-linear measurement system will over-report the acceleration amplitudes of frequencies that occur often as sidebands and intermodulation products in a complex mechanical system, this phenomenon is noted here. The two-times rotor rotational signal at 8.60 Hz is reported at -50.63 dB by the accelerometer but at -63.48 dB by the RLI. This difference means that more than eighty percent of the energy in the signal reported by the accelerometer is sensor system artifact. A similar 9 dB difference is seen at five-times rotor rotational.

Mechanical Origin of Signal	Accelerometer Amplitude	RLI Amplitude
1 X Main Rotor Rotational (4.30 Hz)	not seen	-69.70 dB
2 X Main Rotor Rotational (8.60 Hz)	-50.63 dB	-63.48 dB
3 X Main Rotor Rotational (12.88 Hz)	not seen	-63.45 dB
4 X Main Rotor Rotational (17.19 Hz)	not seen	-59.78 dB
Main Bevel Gear Rotational (20.09 Hz)	not seen	-61.13 dB
5 X Main Rotor Rotational (21.48 Hz)	-44.54 dB	-53.41 dB
6 X Main Rotor Rotational (25.78 Hz)	not seen	-56.54 dB
Main Pinion Rotational (95.70 Hz)	-28.03 dB	-28.39 dB
Hydraulic Pump Rotational (119.65 Hz)	-33.45 dB	-34.40 dB
Input Pinion Rotational (348.00 Hz)	1.27 dB	.46 dB
Planet Gear Mesh (979.62 Hz)	4.90 dB	5.29 dB
TTO Bevel Gear Mesh (1,507.27 Hz)	5.92 dB	4.97 dB
Main Bevel Gear Mesh (2,009.69 Hz)	9.90 dB	7.49 dB
Input Pinion Gear Mesh (7,656.00 Hz)	8.38 dB	11.19 dB

Measurement Results (H-53E Main Gearbox)

Similar to the SH-60 Main Gearbox data set presented above, the H-53E data is presented in figure 2 which shows the 16-second sample time measurements of the accelerometer above the simultaneous measurement of the RLI. In the interest of

brevity, these plots are only shown to 96 Hz, which also displays the range where the RLI measurements have a substantially lower noise floor than the accelerometer measurements. Most of the first ten harmonics of the main rotor rotational speed (2.98 Hz) are clearly present in the RLI signal. Additional low frequency vibrations are also identifiable in the RLI measurements.

The next data presentation for this test condition and location is the table of frequencies and amplitudes that had substantial signal-to-noise ratio and that appeared at least 58 percent of the time. The RLI table shows the first (2.98 Hz), second (5.96), third (8.94 Hz), and seventh (20.91 Hz) harmonics are reported. In addition, the second sun gear rotational (7.7 Hz), second planet rotational (16.22 Hz) and Main Bevel Rotational (29.39 Hz) are reported.

Mechanical Origin of Vibration	Accelerometer Measurement	RLI Measurement
Main Rotor Rotational (2.98 Hz)	not seen	-67.00 dB
2 X Main Rotor Rotational (5.96 Hz)	not seen	-75.93 dB
Second Sun Gear Rotational (7.07 Hz)	not seen	-76.89 dB
3 X Main Rotor Rotational (8.94 Hz)	not seen	-72.02 dB
Second Planetary Rotational (16.22 Hz)	not seen	-66.97 dB
7 X Main Rotor Rotational (20.91 Hz)	not seen	-51.85 dB
First Planetary Rotational (23.89 Hz)	not seen	-68.69 dB
Main Bevel Rotational (29.39 Hz)	-37.12 dB	-41.32 dB
#2 Input Bull Gear (77.56 Hz)	-23.11 dB	-27.02 dB

Conclusion: The RLI system makes non-contact vibration measurements of helicopter gearboxes that are comparable to accelerometer readings in the worst case and considerably better than accelerometer readings in some frequency ranges of interest. The ability to measure vibrations of known mechanical origin within the gearboxes in frequency ranges where accelerometer measurements do not provide measurement capability has been demonstrated.

Acknowledgements: This effort was undertaken for the Naval Air Systems Command Code 4.4.2 and the Naval Air Warfare Center Aircraft Division, Patuxent River, Maryland. We would like to thank Mr. Mark Agnello and Mr. Mike Richman for their support. The measurements were made at the Naval Aviation Depot, North Island Detachment, Pensacola, Florida and we would like to thank Mr. Ralph Horne.

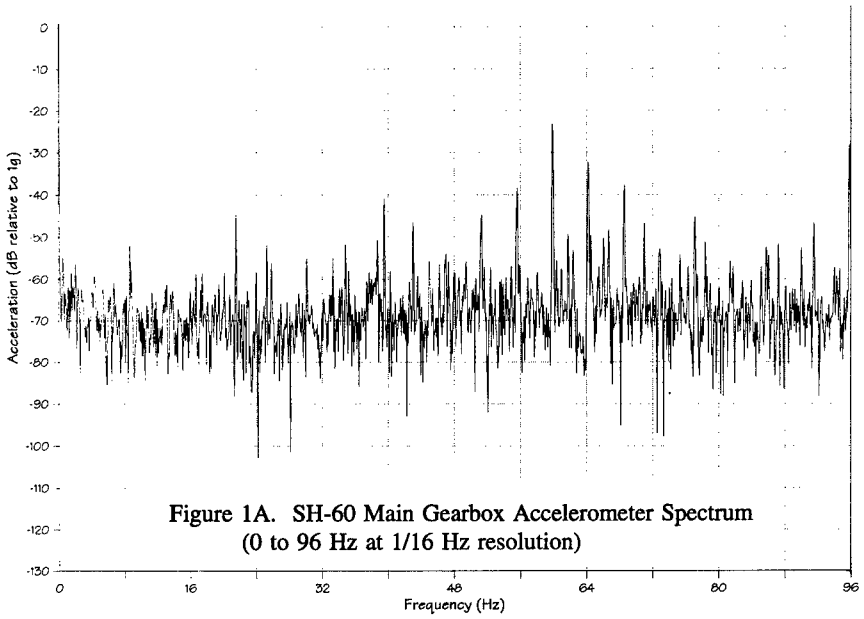


Figure 1A. SH-60 Main Gearbox Accelerometer Spectrum
(0 to 96 Hz at 1/16 Hz resolution)

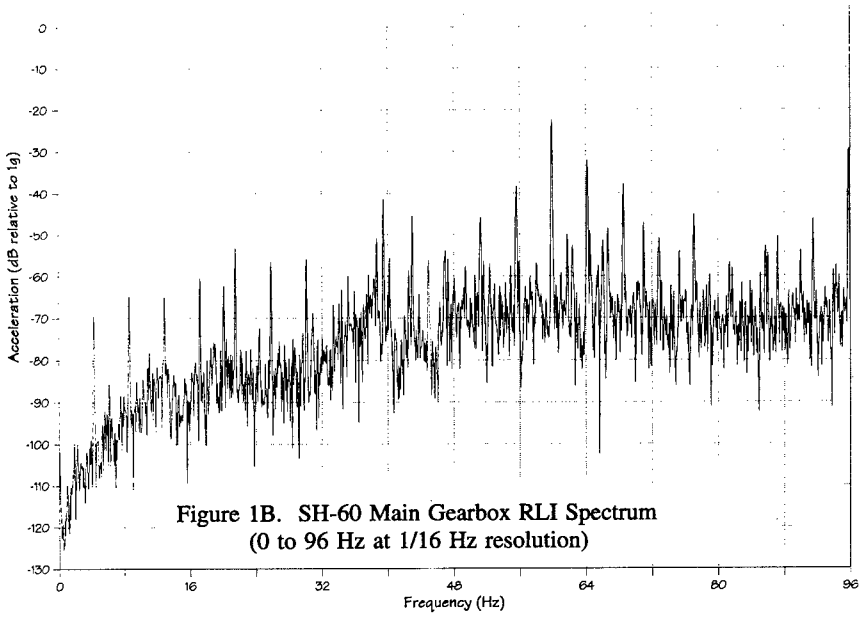


Figure 1B. SH-60 Main Gearbox RLI Spectrum
(0 to 96 Hz at 1/16 Hz resolution)

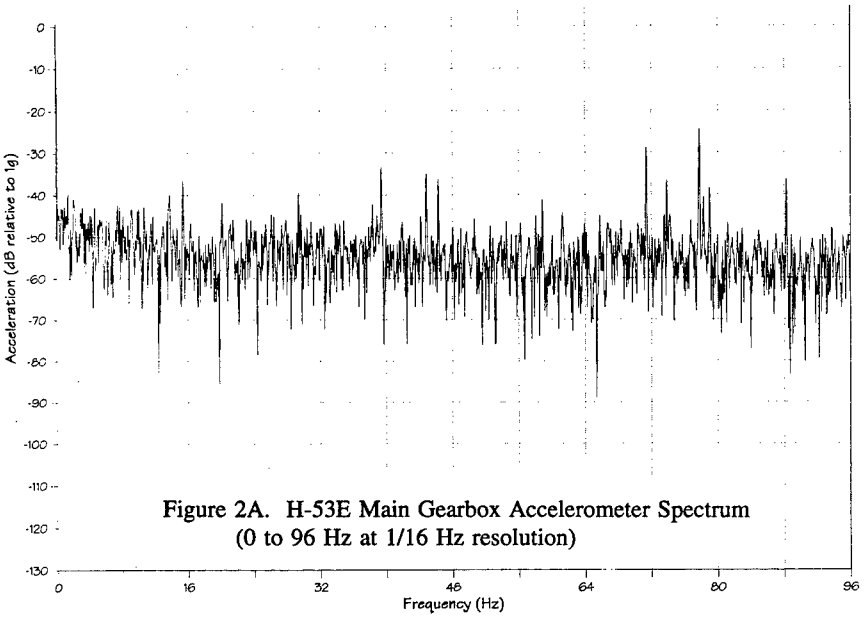


Figure 2A. H-53E Main Gearbox Accelerometer Spectrum
(0 to 96 Hz at 1/16 Hz resolution)

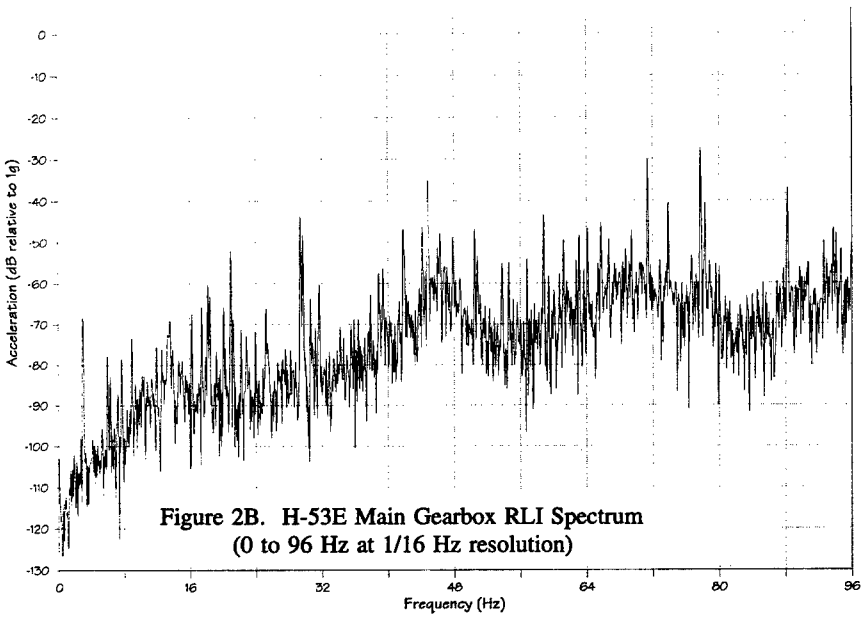


Figure 2B. H-53E Main Gearbox RLI Spectrum
(0 to 96 Hz at 1/16 Hz resolution)

A PROGNOSTIC MONITOR FOR MICROELECTRONICS SYSTEMS RELIABILITY

Wilmar W. Sifre, Steven L. Drager and Martin J. Walter
Rome Laboratory/ERDD
525 Brooks Rd.
Rome, NY 13441-4505

Abstract: An innovative use of test structures for increasing the maintainability, reliability and availability of electronic systems is presented. The proposed use scenario provides insight into the real-time status of the reliability of the microelectronic components of the system through a prognostic reliability monitoring vehicle. Specifically, this technique provides the user with a library of prognostic cells and a test methodology which may be utilized to predict an impending operational failure of electronic components, modules, and systems due to the failure mechanisms of hot carrier degradation and oxide breakdown.

Key Words: Hot carrier degradation; microelectronics; oxide breakdown; prognostics; reliability

Introduction: The ability to identify components, modules and subsystems which are likely to fail will significantly enhance the availability of the system. This identification may lead to the replacement of the most vulnerable systems with systems that are sure to survive. This means that critical systems will function properly throughout their service, therefore reducing unscheduled maintenance. This is of particular interest to the Air Force and DoD where electronic systems such as, radar, and jammers are critical to the survival of the aircrew and aircraft in combat situations.

Prognostic knowledge reduces the overall maintenance costs by moving from a preventive to a predictive maintenance scheme. The work presented here describes an approach which provides the necessary insight into the real-time status of the reliability of the electronic components, modules and subsystems. This approach utilizes the new idea of reliability prognostic monitor cells and standard boundary scan found in many current electronic systems.

The reliability prognostic monitor cells are designed to be self contained predictors of impending circuit failure [1,2]. In this paper two such monitors are described, one for hot carrier degradation and another for oxide breakdown (also known as Time Dependent Dielectric Breakdown, TDDB). This paper also addresses how these prognostic cells can be integrated with the IEEE 1149.1 Standard Test Access Port and Boundary Scan Architecture and possible application scenarios for an electronic system.

The prognostic cell capability is based on a library of monitor cells which have been designed to provide the diagnostic foundation for monitoring and assessing the reliability degradation experienced by the Application Specific Integrated Circuits (ASICs). Each of

the prognostic reliability monitor cells in the library has been designed to be capable of performing accelerated reliability tests, similar to those conducted at the wafer level for process monitoring. The prognostic monitors employ accelerated stress conditions to increase their rate of degradation as compared to the degradation of the functional circuit components. This acceleration provides assurance that the monitor will fail before the functional circuit. Including the monitors as part of the integrated circuit makes certain that the environmental conditions experienced by the integrated circuit (including fabrication, packaging, burn-in and operational conditions) are experienced by the monitor as well. This assures that variations in temperature and voltage for the monitor and the functional circuit elements will be the same. This is important as traditional methods of lifetime prediction are based on the conditions experienced by the integrated circuit up to insertion into a system. The proposed techniques take into consideration any impact the operating environment may have on reducing the lifetime of the integrated circuits.

Incorporating the prognostic cells on-chip produces several requirements which are: the size of the prognostic monitors and any associated circuitry must be as small as possible; the power consumption of the monitor circuits must be minimal so as not to impact the operation of the integrated circuit; and the available power to the prognostic circuitry is limited to the integrated circuit supply voltage. With these requirements in mind, the cells were designed so that each monitor cell is composed of three sections: circuitry which introduces a stress to a test structure, a test structure which is under stress, and circuitry to monitor the degradation which has occurred. This configuration provides real-time degradation information to the system maintainer for use in proactive maintenance, as opposed to predicted lifetime values used for preventive maintenance. This methodology also provides the unique capability to reveal any impact to the system reliability resulting from changes due to the thermal, mechanical and electrical stresses which the system is subjected to during assembly, integration and operation.

Prognostic Theory: The theory behind the operation of the prognostic cells may best be shown through the example of oxide breakdown. Figure 1 depicts the expected oxide lifetimes at varying electric fields for an intrinsic (diamond points) and a defective (square points) oxide. As may be noted from this figure, for a given electric field, the defective oxide will reach failure before that of the intrinsic oxide. For example, at a 5.0 volt supply voltage (line A in Figure 1), a 10nm oxide experiences a field of 5MV/cm. At this field, the intrinsic oxide has a lifetime around 6.24E+21 seconds, while the defective oxide has an expected lifetime of 6.24E+15 seconds. One may also see from this chart that an increase in the field across the oxide reduces the expected lifetime of the oxide whether it is intrinsic or defective. Based on this theory, if the prognostic monitors are placed at a higher operating field than the oxides of the functional circuitry on the chip, one may expect to see failure of the monitor cell before the actual integrated circuit in which the prognostic monitor resides.

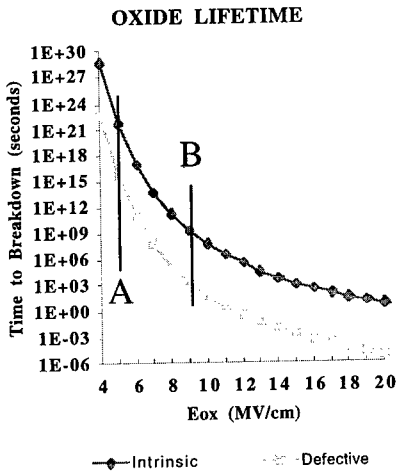


Figure 1. Prognostic Theory.

For example, an oxide prognostic monitor operating at a 9.0 volt supply voltage (point B in figure 1) and having a 10nm oxide thickness would be stressed at 9.0MV/cm. The intrinsic life of that oxide would be 2.76E+09 seconds, while the defective lifetime of that oxide would only be 2.76E+03 seconds. Both of the oxide lifetimes for this higher supply voltage are lower than the expected oxide lifetimes for the functional circuitry operating at a 5 volt supply voltage. Thus, even for the worst case scenario, that is where the prognostic cell oxide is intrinsic and the oxides of the functional circuitry are defective, the prognostic monitor will fail earlier than the functional oxides.

Prognostic Cells: Figure 2 shows a block diagram of the prognostic monitor cells and the test methodology proposed by this research. By incorporating the prognostic cells into the IEEE Boundary Scan Test Access Port (TAP), the maintainer is provided with an easy means by which to read the status of the cells. The major components of this methodology are shown in the block diagram. They are the standard Test Access Port controller, the test circuitry for the prognostic cells, the static and dynamic hot carrier prognostic cell blocks and the oxide breakdown prognostic cell block.

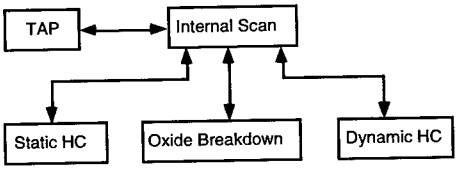


Figure 2. Prognostic Cell Functional Diagram.

The TAP controller integrates the prognostic process with a system level maintenance test bus. It provides the maintainer with a means to assess the current state of electronic circuit reliability. A methodology to assess the state of the prognostic cells through the TAP has been developed and simulated. This methodology provides an easy way for the prognostic circuitry to reside along with the functional circuitry without interfering with the normal circuit operation.

The Oxide Breakdown Cell is shown in Figure 3. This circuit is composed of a gated clock buffer (transistors P1, P2, P3, P4, N1, N2 and N3), a well bias generator (transistors P5, N4 and N5), and gate oxide failure detector (transistor P6, P7, N6, N7 and N8).

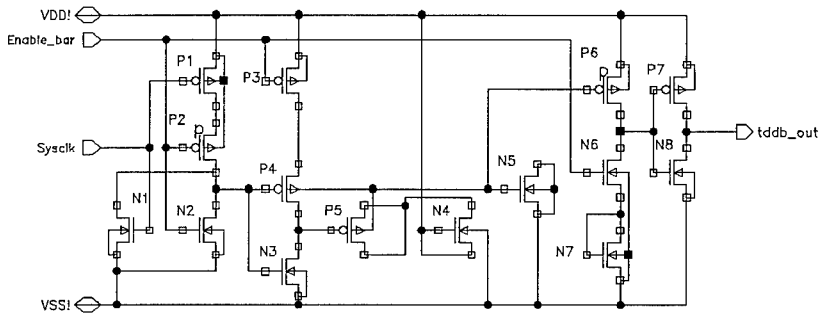


Figure 3. Oxide Breakdown Cell.

The clock buffer connects the gate of transistor P5 alternately between the power rail and ground. When the clock buffer is in the low state, the channel of P5 is charged to a voltage one N-channel threshold below the power rail. The well bias generator circuitry transfers charge from the power rail to the well, which causes the well potential to increase beyond the power rail potential. When the clock buffer is switched to a high state, the channel charge on P5 is driven into the well through the source/drain junctions. The charge injected into the well is shared with the gate oxide capacitor formed by N5. The P5 gate oxide is also being stressed, but at the clock duty cycle. As the clock driver switches alternately high and low, the potential on the well and the gate of N5 rises. Simulation results show that the well potential rises to about 9.0 volts within 100 μs, see Figure 4a. This voltage limit is determined by the threshold of N4, P5 and the forward junction drop of the source/drain of P5. The maximum well bias will be determined by the final manufactured threshold of the N-channel and P-channel transistors.

Enable_bar is an active low signal that enables the well pump circuitry and disables the output sensing. P3 serves as an N-well isolator (P3 and P6 share the same well, which is isolated from the pumped well) so that the charge pumped well can be disconnected from the power supply rail during failure testing. Finally, Sysclk is the pulse clock source which may be almost any frequency on the order of 5.0 MHz or greater. The clock can

be derived from an existing ASIC clock source or it can be supplied by a ring oscillator contained within the prognostic cells.

To test the state of the cell, Enable_bar is set to a logic high value. This disconnects P4 from the power rail and the charged well slowly discharges through the reversed biased well junction, the source drain junctions and through the Fowler-Nordheim current in the gate oxide. Gate oxide failure for transistors N5 and P5 is detected via transistor P6. This is accomplished by connecting P6 in the following manner: the gate is connected to the pumped n-well, the source to the supply voltage and the drain to the sensing circuitry. As long as the test capacitors (N5 and P5) have no significant leakage, then P6 will not conduct, because the gate-source potential is greater than zero. If the gate oxide in either N5 or P5 has a breakdown, then the P6 drain voltage will be dragged to a potential that is determined by the threshold of N4 plus a forward diode drop. The result is that P6 will turn ON causing TDDB_OUT to go high when the gate voltage drops a threshold below the 5.0 volt rail. Currently, there is no way to distinguish between a failure in N5 or P5.

Since there are several normal leakage paths in this circuit, the charge condition of the well is a dynamic state when the clock driver is disabled. In order to obtain an accurate assessment of the condition of the gate oxides under test, it is necessary to time the testing of the output properly. If the output is monitored too quickly after disabling the clock driver, it may be possible to overlook a gate oxide failure that is high resistance on the order of 500KΩ. On the other hand, a long delay may lead to erroneously assigning failure to a good test structure. An assumption was made that there will be very few cases where a failed gate oxide will exhibit a leakage resistance greater than about 500KΩ. This was used as a basis for determining the minimum measurement delay for detecting a high resistance oxide failure.

Simulations were run to determine whether a minimum delay time is required to guarantee that the "leaky" capacitor case is detectable. These simulations indicate that a minimum delay should be in the range of 50μs to unambiguously detect this case. In addition, it is necessary to recognize that the well bias will decay when the clock driver is disabled during testing. In order that a good gate oxide capacitor not be logged as defective, it is necessary to restrict measurement delay time to less than 1.0ms. This avoids junction leakage and subthreshold leakage caused bias discharge.

This circuit has been simulated, as shown in Figure 4. The cell was subjected to stress during the intervals between 0 to 25 μs and 50 to 75 μs while the oxide integrity was evaluated during 25 to 50 μs and 75 to 100 μs. The simulations of failed oxides were performed using a 500KΩ resistor to short the oxide under stress to ground, while the simulations of good oxides have no resistor in place. Figure 4a shows the oxide prognostic monitor simulation results when the oxide is intact. The well bias approaches 9.0 volts, while the output voltage remains at logic high. Figure 4b displays the simulation results for the case of a failed or leaky oxide. The well bias is seen to decay during the early portion of the test cycle and the output voltage pulses high for the early part of the test cycle, but is low during the time that the output voltage will be evaluated.

From this SPICE output, one may easily see that the operation of the oxide breakdown prognostic cell provides distinctive output for both the good and bad oxide.

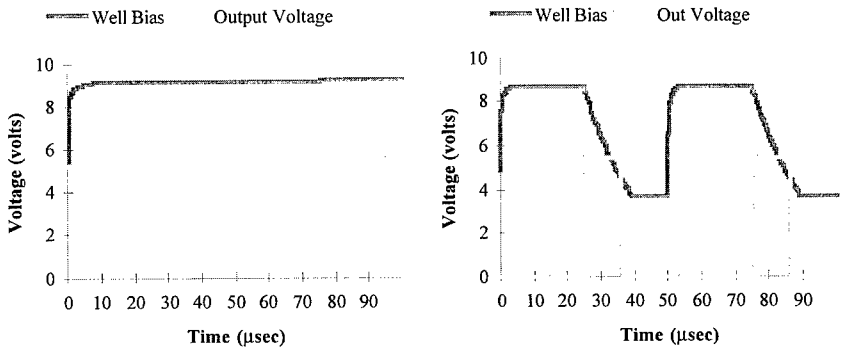


Figure 4. Oxide Breakdown Cell Operation. 4a. The left picture shows a good oxide. 4b. The right picture shows a shorted oxide.

The Dynamic Hot Carrier Cell is composed of two identical voltage controlled oscillators, as seen in Figure 5. In the proposed configuration, one oscillator is running whenever power is applied to the circuit. This oscillator is the device under test for this monitor, as its transistors experience degradation conditions whenever the IC is powered. The second oscillator is used as a baseline for the degradation, meaning that it is only turned on during a measurement cycle. Hot carrier degradation is characterized by a decrease in the mobility and an increase in the threshold voltage of the nmos transistors. As a result, the rise, fall and delay times of the voltage controlled oscillator are increased. This increase in the delay time is directly observed as a decrease in the oscillator frequency. The amount of degradation which has occurred may be assessed by comparing the frequency of oscillation of the two oscillators.

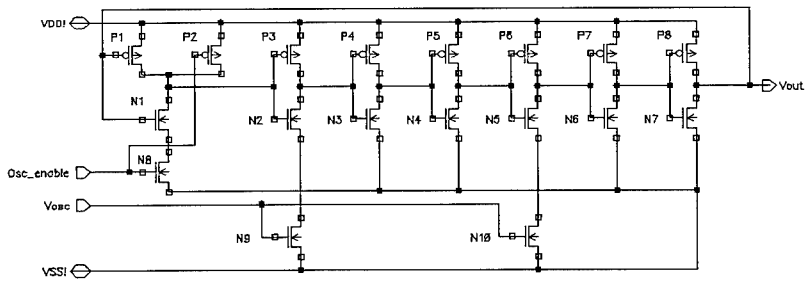


Figure 5. Voltage Controlled Oscillator.

The Static Hot Carrier Cell is shown in Figure 6. The primary goal in this design was the ability to resolve relative differences in the saturation current between a test and reference transistor. The design is composed of a differential amplifier with one transistor under stress (DUT) and a second transistor unstressed (REF). The stress of this prognostic cell is performed under worst case hot carrier bias (Vds=Vdd, Vgs=Vds/2, Vs=0, Vsub=0). This cell will monitor changes in the mobility which are detected by a change in the drain current. The drain current is amplified by the differential amplifier and then fed into a current controlled oscillator. As in the dynamic hot carrier cell, degradation is assessed by comparing the frequency of oscillation between the DUT and REF transistors.

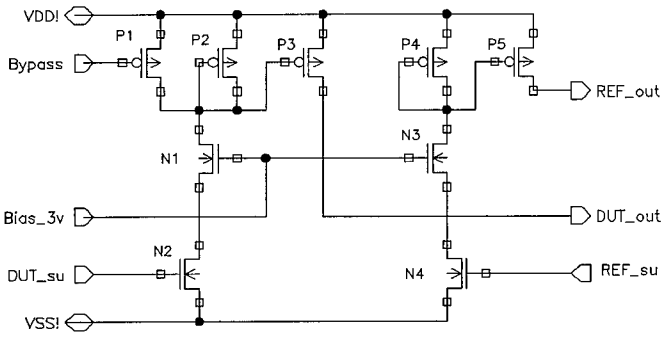


Figure 6. Static Hot Carrier Cell.

Even though every effort was made to produce identical transistors for both DUT and REF, it must be recognized that there will be some differences between them due to process variations. It must also be recognized that the measurement cycle results in a small amount of hot carrier degradation to the transistors. However, both of the transistors will be exposed to identical amounts of degradation. Therefore, because of the process variations as well as the small degradation incurred during measurement, it is important to make an initial current measurement on both transistors before any stress is applied. Finally, in order to compensate for the possible existence of mobile ions the gate bias, Bias_3v, remains applied to the reference transistor during the stress cycles on the test transistor. Any threshold changes due to mobile ions will then effect both transistors in the same way.

The test transistors are indicated in Figure 6 as N1 (DUT) and N3 (REF) and are interdigitated pairs of transistors. Interdigitated layout is used in all parts of this design where it is desirable to have good matched characteristics between the stressed transistor and the unstressed transistor as well as their respective current monitoring circuits. A similar interdigitated structure is used for the current mirrors which are used for the current measuring circuits. P2 and P3 form the current mirror for the DUT, while P4 and P5 form the current mirror for the REF. The sources of each of these transistors connect to the drain of a disconnect transistor (N2, for DUT, and N4 for REF). This provides a means for interrupting the current flow. Finally, P1 is used to bypass the current mirror

in the drain of the stressed transistor during the stress cycles by placing the signal Bypass at ground.

Stress is applied to only the DUT transistor by placing the control ports Bypass and REF_su to ground and the control port DUT_su to the supply voltage. The gate bias for both transistors is set to Vdd/2 for stress. Measurement of the drain current of the test transistors is accomplished by setting control ports Bypass, DUT_su and REF_su to the supply voltage. The gate bias remains set to Vdd/2 because it does not matter what the gate bias is for measurement of the drain current provided that the gate bias is greater than the threshold voltage. The drain current measurement is performed by alternately connecting REF_out and DUT_out to the current controlled oscillator.

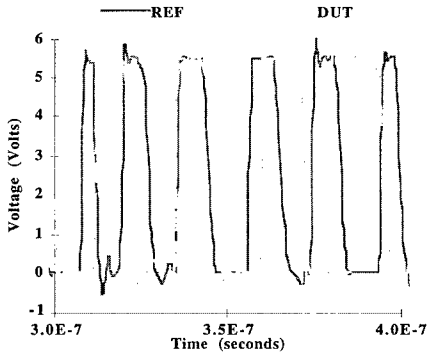


Figure 7. Static Hot Carrier Cell Simulations.

Figure 7 depicts the SPICE simulation results from the static hot carrier cell. This figure shows the output of the current controlled oscillator for both the DUT and REF transistors. One may see that the DUT transistor has a lower frequency than the REF transistor, thus showing hot carrier degradation has occurred. The simulations were performed using a degraded SPICE model for the DUT transistor. The SPICE model parameters were obtained from measurements on both hot carrier stressed and unstressed N-channel transistors.

Applications of Prognostic Techniques: This work has demonstrated that a diagnostic capability will be provided to the system maintainer by having the prognostic monitor cell design included in the ANSI/IEEE Standard 1149.1 internal scan control chain. This will provide a common built-in-test architecture to enable the tracking of end of life failure mechanisms such as oxide breakdown and hot carrier degradation. When these prognostic reliability cells are added into ASIC systems, visibility into the physical environment and access to the reliability characteristics of the host system are established.

In the field, prognostic monitors provide a vehicle for assuring that systems used for highly critical applications are not going to fail. Using the Boundary Scan Test Access

Port, the system maintainer is able to query the state of the prognostic monitor, determining the likelihood that the functional blocks of the circuit will continue to function and allow for mission success.

Another plausible scenario for use of these prognostic cells is the evaluation of new circuit packaging schemes. In this scope simple specialized reliability test chips would be created and packaged for monitoring and evaluating the new packaging environments. The results would provide the manufacturer with feedback on the reliability implications of the proposed packages and system designers with knowledge useful before committing the design to fabrication.

Conclusions: The use of prognostic techniques will improve the ease of use of reliable electronic systems; reduce the cost associated with the extended use of systems; and reduce the effects of electronic failure on operational performance. Through predictive maintenance, costs are reduced while the availability, reliability and maintainability of the systems are increased. Finally, the maintainer is provided with a prognostic capability, through the use of a maintenance test bus, which will provide a go/no go status on the state of the reliability of the microelectronics in the system.

This paper has described prognostic reliability monitor cells for oxide breakdown and hot carrier degradation as well as a measurement methodology for these failure mechanisms. Prototype prognostic monitor cells for oxide breakdown and hot carrier degradation (both static and dynamic) have been designed and simulated. Extensive SPICE simulations demonstrating cell functionality and the ability to monitor degradation have been performed. This implementation should allow for the measurement of frequencies up to 100MHz with 16-bit resolution.

This work is still in its infancy. The prognostic cell performance must be verified by incorporation into an ASIC design. An in-house Arithmetic Logic Unit (ALU) design has been chosen for this next step. Once fabricated, the ASIC will be placed on an operational test and the prognostic cell information correlated to the ASIC failures. The testing will provide feedback on the appropriate degradation threshold and define the relationship between the degradation or failure of the prognostic monitors and the functional circuitry.

Acknowledgments: The authors gratefully wish to acknowledge the continued support received from Vance Tyree, University of Southern California Information Sciences Institute; and Daniel Burns, Rome Laboratory who helped with the original inception for this research. We also wish to thank Mark Pronobis, Rome Laboratory, who helped significantly with the IEEE 1149.1 Standard Test Access Port and Boundary Scan Architecture; and Robert Hillman, Rome Laboratory, who provided support with the test methodology.

References:

[1] Vance Tyree and Steven Drager, "Prognostic Reliability Monitors," 1994 Dual-Use Technologies and Applications Conference Vol II p152.

[2] Vance Tyree, "Self Stressing Test Structure Cells," RL-TR-95-18.

HIGHLY EFFECTIVE REAL TIME ADAPTIVE HOLOGRAPHIC INTERFEROMETRY AND APPLICATION TO NONDESTRUCTIVE TESTING IN MACHINERY

George E. Dovgalenko, Anatoli V. Kniazkov, Yuri I. Onischenko, Greg J. Salamo
226 Physics Building, The University of Arkansas, Fayetteville, AR 72701

Abstract: This paper provides a overview of developments and progress in interferometric monitoring of diffusely reflective objects and potential applications to machinery. An assessment is made of current interferometric capabilities in monitoring of vibration modes, residual stress, and defects. The new highly effective dynamic hologram technique of adaptive interferometry monitoring of diffusely reflective objects are described. Application of this method for the visualization of displacements by vibration of diffusely reflecting object has been performed. Potential application of this technique to measure the residual stress in a welded joint and defect monitoring is discussed.

Key Words: adaptive interferometry of diffusely reflective objects; defects; high interference contrast; non contact monitoring in real-time displacements, residual stress.

INTRODUCTION: Application of real time interferometry to nondestructive testing of machinery is one of the challenging problem of contemporary engineering. The basic principle of interferometric methods is based on non-contact sensing of displacements which leads to visualization of defects vibration modes and stress. In machinery almost all monitored objects have diffusely reflective surfaces (non mirror-like). Unfortunately a small interference contrast of diffusely reflected objects using contemporary hologram interferometry technique inhibits applications of these methods to nondestructive testing problems of machinery. The most contemporary potentially powerful method of electro-optic holography, also known as electronic holography EH, is the method based on recent advances in hologram interferometry, speckle metrology, and computer technology is not applicable in real time to vibrating object measurements when object vibrates and moves simultaneously or vibration modes fluctuate in time. In an ideal vibrating case (no object movements and no fluctuations in time) EH method based on capture of series of interferograms using CCD camera , computer storage and decoding after some delay time. Unfortunately EH method produced low contrast interferogram of diffusely reflected objects (Fig.1). For these objects however the interferograms need to be seen in real time with high contrast. The real time objects monitoring which based on surface displacements measurements under high energy influence such as laser shout, ballistic problem in military etc. needs reliable non-contact real time metrology. To our knowledge no such systems for diffusely reflected objects monitoring exist. The new technique of highly effective dynamic grating has been proposed for read and readout of highly contrast adaptive holographic interferogram of diffusely reflective objects in real time. The advantage of this technique is that it provides continuous, high contrast

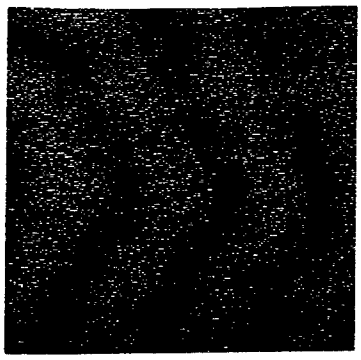


Fig.1. Typical electronic holography fringe pattern of diffusely reflected object.

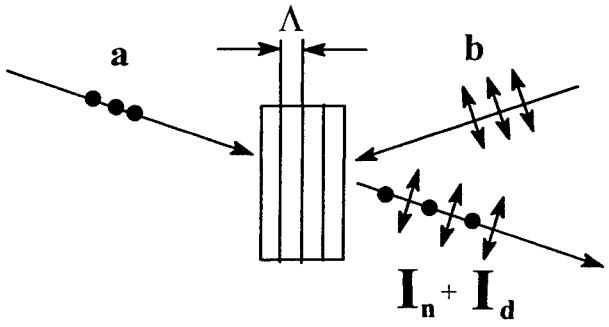


Fig. 2 Anisotropic selfdiffraction of reflection type grating
Λ - grating space, I_n - nondiffracted intensity beam, I_d - diffracted intensity beam, **a** and **b** - vector amplitudes of initial optical waves.

diffusely reflective objects monitoring in real-time and has potential to be a portable device. This technique uses low power He-Ne laser and a photorefractive storage crystal. This leads to reliable and cost effective solution for nondestructive monitoring of machinery. Application of this method for the visualization of displacements by vibration of a diffusely reflecting object has been performed. Some observation on technological gaps and problems to be solved such as measurement the residual stress in welded joint and defects monitoring in metal are discussed.

REGISTRATION SENSOR PHILOSOPHIES: The ideal optic registration element ORE for real time adaptive interferometry should have high diffraction efficiency, high ratio I_{signal}/I_{noise} (high interference contrast), high resolution, low time response, high photosensitivity, no pretreatment before writing, erasing, reading, storage; unlimited number of cycles without optical degradation, easy to operate, small size, reliable in harmful environmental condition, low cost.
In the last two decades the progress in optical engineering was achieved in research and development of new optical materials for ORE based on reversible (reusable) photorefractive crystals and polymers [1]. Unfortunately industrial applications of these materials have been inhibited because non-fulfillment of ideal requirements listed above. The most sensitive photorefractive crystals sillenite family have 10^{-3} J/cm^2 holographic sensitivity compared with silverhalide photomaterials[1] and smallest response time among all photorefractivity but unfortunately have small diffraction efficiency in the range of 0.1% - 1% [1], [2], [3]. In practical using it causes application of high electrical fields (6.000-10.000 V/cm), decreasing the ratio I_{signal}/I_{noise}, creates optical damages in crystal. This is impractical for optical devices developments [4] and dramatically inhibits applications of ORE for noncontact monitoring. The highest diffraction efficiency photorefractive crystals such as $LiNbO_3$, $BaTiO_3$ have long response time, and small holographic sensitivity. It causes application to long time storage and does not applicable to on line monitoring. The shortest response time bacteriorhodopsine polymer films have small diffraction efficiency in the range of 1 % - 6 % [5]. The thermoplastic materials needs pretreatment like as heating, electrical field application (kilovolts) and have small number of cycles [6]. All these materials have small space resolution which causes a small contrast in interferogram of diffusely reflective objects. One can see these materials can not fulfill to ideal ORE requirements and not reliable for development of optical devices for real-time adaptive interferometry of diffusely reflective objects. The aim of this paper is to develop and design a reliable ORE usable for optical applications.

THEORY: Using tensor-vector algorithm [6] in geometry of two counterpropagated waves Fig.2 for photorefractive crystals symmetry 23 (BSO, BGO, BTO) amplitude of reflected wave is:

$$\mathbf{a}(z_0) = [\mathbf{M}\mathbf{a}^0 + (\mathbf{M}_a\mathbf{a}^0\mathbf{M}_b\mathbf{b}^{0*})n_0^3 z_0 k_0 \mathrm{FR}\mathbf{M}_b\mathbf{b}^0]\exp(ik_0 n_0 z_0) \quad (1)$$

where:

$$M_{b>a} = \begin{vmatrix} \cos\varphi & \pm\sin\varphi & 0 \\ \mp\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{vmatrix} \text{-are optical activity matrices} \qquad (2)$$

$\mathbf{a}^0$, $\mathbf{b}^0$ - are initial amplitudes of optical waves, $\varphi = Gn_0^3 z_0 k_0/2$ -is optical activity rotation angle, $G=G_0 I$ - is gyration pseudotensor, I - is unit matrix, G_0- is gyration parameter, $n_0=(\varepsilon^0)^{1/2}$ - is crystal refraction index, z_0- is crystal thickness, $k_0=2\pi/\lambda_0$ - is vacuum wave number, λ_0 - is vacuum wave length,

$$R= r_{14} \begin{vmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{vmatrix} \qquad (3)$$

R- is tensor of linear electrooptic effect in crystallographic axes$(1\,\bar{1}\,0),(110),(100)$,
$F= E_d(1 + E_d/E_q)I_0$, $E_d = 2\pi KT/\Lambda e$ - is diffuse field in crystal, K- is Bolzman's constant, T - is temperature, e - is electron charge, $E_q = 2eN_A \Lambda/\varepsilon_r$ - is space charge field, N_A - is trap density in the crystal volume ε_r - is static dielectric permittivity, $I_0 =[M_a\mathbf{a}^0]^2 + [M_b\mathbf{b}^0]^2$ -is counterpropagated waves intensity. When grating are formed by orthogonal waves $\mathbf{a}^0 = (a_1,0,0)$, $\mathbf{b}^0 = (0,b_2,0)$
parameter $M_a\mathbf{a}^0\, M_b\mathbf{b}^{0*} \neq 0$ in equation (1) and vector outgoing wave is:

$$\mathbf{a}(z_0) = \mathbf{a}_L + \mathbf{a}_{NL} \qquad (4)$$

where $\mathbf{a}_L = M_a\mathbf{a}^0 \exp(ik_0n_0z_0)$ - is nondiffracted vector component of outgoing wave, $\mathbf{a}_{NL} = [(M_a\mathbf{a}^0 M_b\mathbf{b}^{0*})\beta M_b\mathbf{b}^0] \exp(ik_0n_0z_0)$ is diffracted component of outgoing wave, $I_n = (\mathbf{a}_L\mathbf{a}_L^*)$ is intensity of nondiffracted wave, $I_d = (\mathbf{a}_{NL}\mathbf{a}_{NL}^*)$ is intensity of diffracted wave, where $\beta = r_{14}n_0^3k_0z_0F$. Analyzing (1) and (2) one can obtain $\mathbf{a}_L \perp \mathbf{a}_{NL}$ That means that polarization transformation of initial linear polarization waves is present. In this case the intensities of outgoing waves are:

$$\begin{aligned} I_a(z_0) &= \mathbf{a}(z_0)\,\mathbf{a}^*(z_0) = |\mathbf{a}^0|^2(1 + [\beta\,|\mathbf{b}^0|^2\sin2\varphi]^2) \\ I_b(0) &= \mathbf{b}(0)\,\mathbf{b}^*(0) = |\mathbf{b}^0|^2(1 - [\beta|\mathbf{a}^0|^2\sin2\varphi]^2) \end{aligned} \qquad (5)$$

If we introduce the diffraction efficiency as: $\mu = (|\,\mathbf{a}(z_0)|^2 - |\mathbf{a}^0|^2)\,|\mathbf{b}^0|^{-2}$
we can obtain

$$\mu = (\beta|\mathbf{b}^0|^2\sin2\varphi)^2|\mathbf{a}^0|^2\,|\mathbf{b}^0|^{-2} \qquad (6)$$

EXPERIMENT:

Set up of experiment is presented on Fig.3. Using equation (2) for amplitude $\mathbf{a}(z_0)$ wave coming through linear polarizer $P(\alpha)$:

$$\mathbf{a}_{out} = P(\alpha)\,\mathbf{a}(z_0) \tag{7}$$

where $P(\alpha) = T(\alpha)\,P(0)\,T(\alpha)^{-1}$, $P(0)$ - is matrix of linear non- ideal polarizer

$$P(0) = \begin{vmatrix} K_1 & 0 & 0 \\ 0 & K_2 & 0 \\ 0 & 0 & 0 \end{vmatrix} \tag{8}$$

where K_1 and K_2 are major transmitting axis coefficients, $T(\alpha)$ is rotated matrix

$$T(\alpha) = \begin{vmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{vmatrix} \tag{9}$$

Matrix $P(\alpha)$ is calculated as

$$P(\alpha) = \begin{vmatrix} K_1\cos^2\alpha + K_2\sin^2\alpha & (K_1 - K_2)\sin\alpha\cos\alpha & 0 \\ (K_1 - K_2)\sin\alpha\cos\alpha & K_1\sin^2\alpha + K_2\cos^2\alpha & 0 \\ 0 & 0 & 0 \end{vmatrix} \tag{10}$$

Matrix $P(\alpha)$ can be presented as $P(\phi) = P_1 + P_2$
where $\alpha = \phi + \pi/2$ is optimized parameter

$$P_1 = \begin{vmatrix} \sin^2\phi & -\sin\phi\cos\phi & 0 \\ -\sin\phi\cos\phi & \cos^2\phi & 0 \\ 0 & 0 & 0 \end{vmatrix} K_1 \tag{11}$$

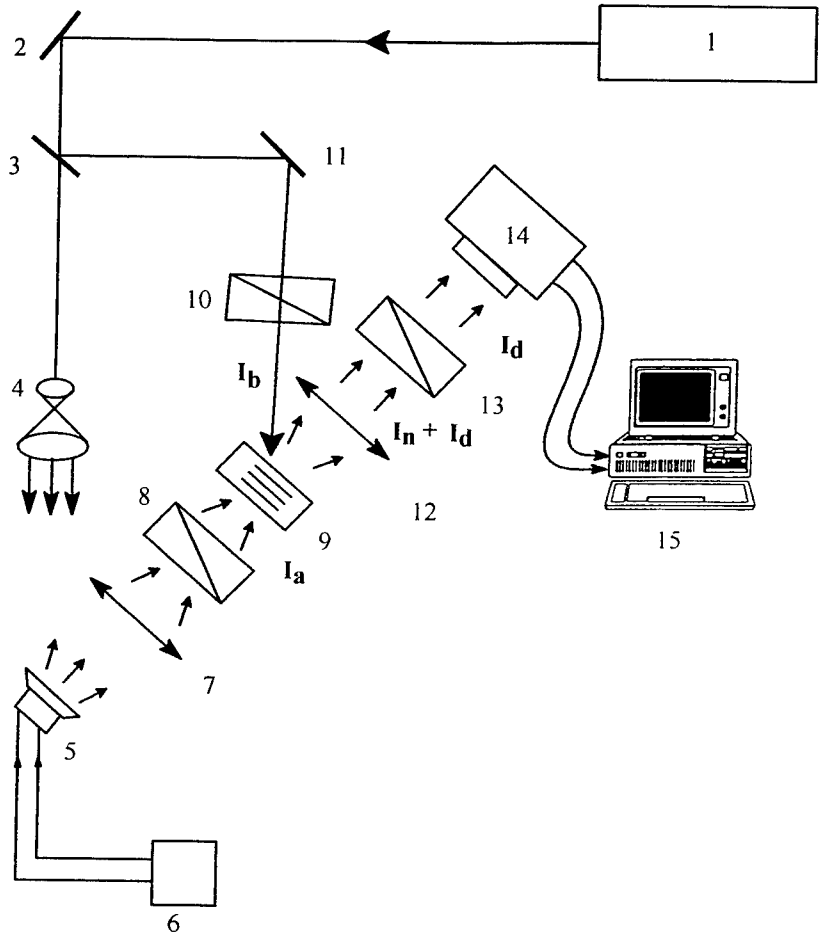


Fig. 3. Sketch of the experimental set-up of adaptive interferometer.
1- CW He-Ne laser; 2, 11- mirrors; 3 - semitransparent mirror; 4 - beam expander; 5 - diffusely reflected vibrated surface (loudspeaker);6 -electric generator; 7, 12 - lenses; 8, 10, 13 - polarizers; 14 - CCD camera; 15 - computer

and

$$P_2 = \begin{vmatrix} \cos^2\phi & \sin\phi\cos\phi & 0 \\ \sin\phi\cos\phi & \sin^2\phi & 0 \\ 0 & 0 & 0 \end{vmatrix} K_2 \qquad (12)$$

Using (11) we have $P_1 M_a \mathbf{a}(z_0) = 0$ and

$$\mathbf{a}_{noise} = P_2 M_a \mathbf{a}(z_0) = a_1 K_2(\cos\phi, \sin\phi, 0) \qquad (13)$$

Equation (11) describes the nondiffracted optical vector to be transmitted through nonideal polarizer and is called as noise optical vector. Diffracted vector signal can be calculated as

$$\mathbf{a}_{dif} = P(\phi)(M_a \mathbf{a}^0 M_b \mathbf{b}^{0*}) n_0^3 z_0 k_0 F R M_b \mathbf{b}^0 \qquad (14)$$

Using (13) and (14) we find out that $\mathbf{a}_{dif} \perp \mathbf{a}_{noise}$
Taking into account that $\mathbf{a}_{out} = \mathbf{a}_{dif} + \mathbf{a}_{noise}$
and

$$I_{out} = (\mathbf{a}_{dif} + \mathbf{a}_{noise})(\mathbf{a}_{dif} + \mathbf{a}_{noise})^* = I_{signal} + I_{noise} \qquad (15)$$

where $I_{signal} = \mathbf{a}_{dif}\mathbf{a}_{dif}^*$ and $I_{noise} = \mathbf{a}_{noise}\mathbf{a}_{noise}^*$
The ratio I_{signal}/I_{noise} is:

$$I_{signal}/I_{noise} = (\beta|\mathbf{b}^0|^2 \sin 2\varphi\, K_1/K_2)^2 \qquad (16)$$

Using perfect polarization calcite prism $(K_1/K_2)^2 = 10^6$ we can increase the ratio I_{signal}/I_{noise} (16). This ratio can be optimized at $2\varphi = \pi/2$. The rotating angle φ depends on product of specific optical activity ρ and crystal thickness z_0

$$\varphi = \rho z_0$$

where $\rho = G n_0^3 k_0/2$. In our experiment for BTO crystal it gives $z_0 = 7.5$ mm at $\lambda = 0.632\mu$m, and $z_0 = 4$mm at $\lambda = 0.514$ μm. For BSO crystal it gives $z_0 = 1.28$ mm at $\lambda = 0.514$ μm. The diffraction efficiency parameter μ **(6)** depends dramatically on trapped density centers N_A (Fig.4). In experiment we have = 67 % for doped and radiated BTO crystal. Using Fig.4 we can estimate $N_A = 5.2*10^{16}$ cm^{-3}. This crystal provides good quality interferogram of diffusely reflected vibrated membrane (Fig.5). A high contrast of interferogram is presented on Fig.6. Interference contrast m of this image in row # 539 is 0.973, where $m = (I_{max} - I_{min})(I_{max} + I_{min})^{-1}$
Proposed adaptive holographic interferometer is able dramatically increase the ratio I_{signal}/I_{noise} of interferogram Fig.1 which was made using EH technique.

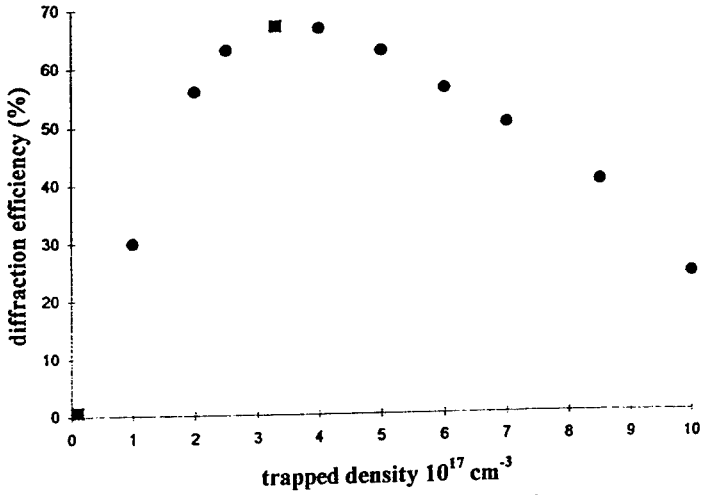


Fig. 4. Diffracted intensity vs trapped density

- **theory**
- **experimental dates**

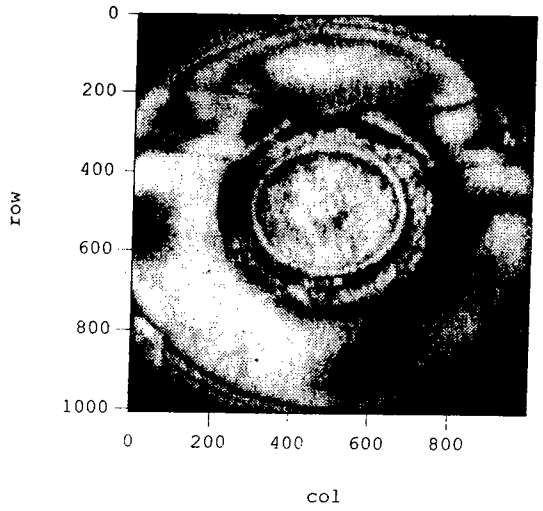


Fig.5. Adaptive real time interferogram of diffusely reflected vibrated membrane

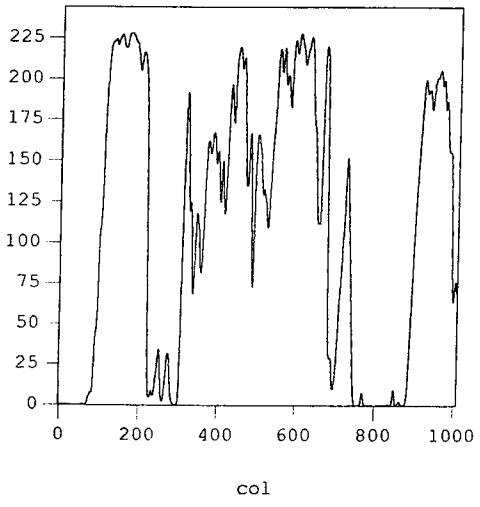


Fig.6. Intensity distribution of row # 539 of interferogram on Fig. 5.

CONCLUSION:

A high effective adaptive grating based on doped and radiated BTO crystal has been made. The high interference contrast of image allows to monitor the displacements of diffusely reflected object such as loudspeaker membrane in the real time. Method has good potential applications for defect monitoring and residual stress testing in machinery on line and can be portable and low cost. The advantage of proposed method in comparison with EH monitoring of diffusely reflective objects has been demonstrated.

REFERENCES:

1. P. Gunter and J.-P. Huignard, *Topics in Applied Physics II*, Vol.62, Springer-Verlag, Germany,1989.
2. D. Webb et al.,"Verification of the standard model of the photorefractive nonlinearity in BSO crystals," *Optics Communications*, Vol. 108, pp.31-36, 1994, North-Holland.
3. G. Uhrich, L. Hesselink, "Temperature, intensity, and field dependence of the absorption coefficient of $Bi_{12}SiO_{20}$," *Optics Letters*, Vol.15, pp.455-475, 1990, USA.
4. M.P. Georges and Ph.C. Lemarie, "Phase-shifting real-time holographic interferometry that uses bismuth silicon oxide crystals," *Applied Optics*, Vol.34, pp.7497-7506, 1995. USA.
5. G. Dovgalenko et al., "Optically induced birefringence in bacteriorhodopsin as an optical limiter," *Applied Physics Letters*, Vol 68 (3), pp.287-289, 1996.
6. U. Schnars, et al., "Digital recording and numeral reconstruction of holograms for nondestructive testing," ed. R.J. Pryputniewicz *Interferometry YII: Applications*, Vol.2545, pp.250-253, 1995, USA.
7. G. Dovgalenko et al., "Optimization algorithm in tensor vector theory of dynamic periodic structures and applications to optical sensors," ed. T. Jannson *Application and Theory of Periodic Structures*, Vol. 2532, pp. 447-475, 1995, USA.
8. N. Kukhtarev et. al., "Reflection holographic grating in [111] cut $Bi_{12}TiO_{20}$ crystal for real time interferometry," *Optics Communications*, Vol. 104, pp. 23-28, 1993, North-Holland.
9. G. Dovgalenko, et al., "Portable Holographic Interferometric Device for Aircraft Inspection," ed. R.J.Pryputniewicz *Interferometry YII: Applications*, Vol. 2545, pp. 265-274, (1995), USA.
10. G. Dovgalenko et al., "Single-beam holographic interferometry and applications," *Technical Digest Photorefractive materials effects and devices*, pp. 295-298, 1995, USA.
11. G. Dovgalenko, et al., "Single-Beam Adaptive Holographic Interferometry -AHI and Small Size Devices for Industrial Inspections," ed. K.G. Harding *Industrial Optical Sensors for Metrology and Inspection*, Vol. 2349, pp.109-113, (1994), USA.

HIGH RELIABILITY ACCELEROMETERS FOR USE IN SHIPBOARD CONDITION ASSESSMENT SYSTEMS

Richard M. Barrett, Jr.
Wilcoxon Research, Incorporated
21 Firstfield Road
Gaithersburg, MD 20878

Abstract: The Pentagon has required that US Naval fleet operation and maintenance be reduced 40% relative to 1992 expenditures. This directive has initiated a reevaluation of condition assessment technology purchased Navy sea systems. Commercial-off-the-shelf (COTS) equipment is envisioned to reduce costs and leverage the latest industrial technology for military purposes. However, Navy condition assessment programs must review the lessons learned in the industrial sector with respect to accelerometer system reliability. The paper below describes accelerometer data quality, reliability and survivability considerations. Mechanical and electrical environmental effects are described with respect to sensor performance. Cabling and mounting issues are discussed as part of the total accelerometer system.

Key Words: Accelerometer; accelerometer cabling; accelerometer mounting; commercial-off-the-shelf (COTS); condition assessment; piezoelectric sensor; vibration analysis

Introduction: Condition assessment programs are being integrated into maritime fleets worldwide. These programs use vibration systems to determine mechanical condition of shipboard machinery, protect equipment from catastrophic damage, and provide diagnostic information for operation and machinery improvement. This technology has been proven to increase machine reliability, vessel safety and lower operational costs.

A key element of any vibration based condition assessment program is the piezoelectric accelerometer. The piezoelectric accelerometer, a device born from early navy test and detection programs, is meeting the new application challenges on the modern fleet. These challenges include high temperatures, steam, electromagnetic interference, physical abuse and cost containment. When selecting shipboard accelerometers data quality, reliability, and long-term survival are paramount to an effective assessment program.

In the past custom accelerometers were designed for multi-general purpose ship-board applications. This was in response to specific military test requirements and vendor driven performance specifications. The result has dramatically increased costs relative to commercial products of similar design and construction. The military now realizes that in many applications COTS (commercial-off-the-shelf) technology can now be leveraged to provide a viable solution at much lower prices. With respect to accelerometers, the military can now take advantage of the lessons learned by manufacturers of industrial sensors systems.

Vibration System Reliability and Data Quality: Vibration system performance depends directly on the reliability and data quality provided by the vibration sensor and signal transmission system. Naval vessels, like most industrial plants, use permanent accelerometers in critical and dangerous areas. The balance of measurements are usually taken with portable equipment. Portable measurements can compromise high frequency data quality due to sensor mounting limitations.

Data quality on the balance of ship systems machinery is being improved through the use of sensors specifically designed for portable measurements and permanent installations of lower cost COTS sensors. Data quality is also improved by preventing environmental effects from interfering with the output signal. These effects include electromagnetic interference, thermal transients, mechanical strains, and distortion from high frequency mechanical noise.

Reliability depends on the sensor linearity, operational performance and stability. Long-term survival of the sensor is directly related to its insusceptibility to both mechanical and electrical failure. Mechanically rugged sensors must withstand chemicals, salt spray, excessive shock and the physical abuse common on shipboard equipment. Electrically robust sensors are required to survive EMI pulses, high temperatures, and thermal cycling.

Sensor system design and selection has traditionally been based on satisfying ambient performance specifications and meeting installed cost targets. However, experience has shown that the physical environment of the application can be the primary determinant of *long-term* system performance and cost effectiveness. Sensors and connection systems are in direct contact with the machine environment. If the materials and construction used in the system cannot withstand the environment they will fail. Sensor system failure will result in costly replacement and reinstallation. Indeed, almost one-third of the investment of a permanently installed vibration system, commercial or military, is allocated to installation labor. It is a matter of simple economics to calculate the cost a "free" warranty replacement adds to the present cost of a sensing channel.

Similarly, the electronics contained with-in the sensor must also withstand the environment. The primary detrimental effects on electronics are high temperatures and electromagnetic events. These factors may cause signal degradation long before failure. Poor data quality can increase costs due to unnecessary analysis and rework. Of course the most significant impact to overall cost would be the failure to provide warning of a developing fault. Losses in terms of both machinery repair costs and personnel safety can easily dwarf the investment of the entire vibration system.
The only way to avoid these catastrophic costs is to do the job right the first time. Therefore it is essential that the system be chosen for long-term environmental survival.

Sensor Survival in Harsh Ship-board Environments: Steam turbine, gas turbine, diesel engine, lubrication and bilge systems all require accelerometer installations built to survive. Installation considerations include component selection, location and cable routing, and system expandability. Material considerations for sensor housings, cabling and connectors are high temperatures, steam, harsh chemicals, lubrication fluids and salt water.

Steam can cause overheating, sealing problems and induce signal overload. High temperatures can significantly shorten the life of electronic circuitry and place constraints on material selection. Harsh chemicals such as cleaning agents can attack many materials. Salt water can produce severe corrosion on many metals including some stainless steals. Ship-board sensors require advanced materials and sealing technology.

Environmentally Rugged Sensor Systems: Materials and Construction: Application differences coupled with instrumentation needs may require the use of a variety of sensors. However, some design factors are common among sensors throughout many environments. Proper material selection is critical to prevent component degradation. Additionally, design construction can positively effect issues such as long-term sealing, ease of installation and removal during machine repair.

Sensor Housings: Materials and Sealing

Sensor housings enclose the sensing element and protect the internal electronics. These enclosures are directly exposed to the worst parts of the machine and must be carefully selected. Type 316L stainless steel provides superior corrosion resistance and has excellent weldability. Originally developed to resist sulfurous acid compounds in sulfite pulp mills, it has become standard for high corrosion environments.[1] Other steels such as Type 303 have exhibited pitting corrosion in salt water environments.

The seals used on the sensor enclosure are absolutely critical. Naval accelerometers require the same sealing and environmental control technology as aerospace electronic hardware. Hermetic seals are required to prevent both liquid and gas entry into the sensor and degradation of the environment manufactured into the sensor. Effective seals are critical to resist entry of lubricating fluids and water. Retention of the "internal environment" of low humidity inert gases is critical for high temperature survival in steam areas.

The military restricts the classification of hermeticity to metal-to-metal and glass-to-metal fusion seals. Welded steel housings and glass sealed connectors provide the sealing levels required for hermeticity. Epoxy and "O"ring seals are improving but cannot be termed hermetic. We recommend that the level of hermeticity be tested on each sensor using helium leak testing. Such testing prevents invisible weld cracks and connector leaks from compromising reliability. No industry standard for sensor leak rates has been written, however Wilcoxon uses 10^{-8} cc/sec of helium as an internal quality assurance requirement.

Connector Configuration: In harsh environments, cables are the first item to be damaged by fire and other operational hazards. Therefore hermetic connector designs are usually preferred over integral cables. The use of a connector ensures sensor integrity even if the cable is destroyed. It is also easier and far more cost effective to replace a damaged cable (or sensor) than a sensor with an integral cable. Modularity lowers vibration system maintenance costs.

Connectors may be located on the top of the sensor (top-connector) or on the side (side-connector). Side-connector sensors generally have a lower profile. This can increase reliability reducing the amount of physical damage to which the connector is physically exposed. Many designs also use a captive bolt through the center of the sensor for ease of mounting and radial connector orientation.

Low profile sensors are also recommended in areas where a top-connector would interfere with machine access during maintenance and cleaning. This prevents service personnel from having to remove the sensor (best case) or "cut" their way through cable (worst case!). The bolt-through design offers the additional advantage of allowing the cable to be easily oriented along the bearing. After the cable is positioned in the desired direction the captive bolt locks it in place. This makes a "cleaner" installation without the use of mounting shims or multiple hole mounting bases. Top exiting sensors are usually less expensive and can be used in areas that are not as confining.

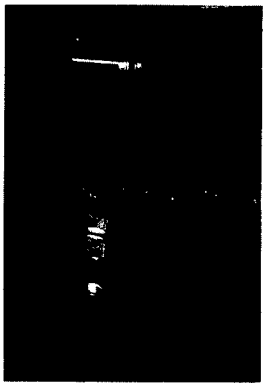


Figure 1: Wilcoxon 786A top-exiting and 787A bolt-thru style Accelerometers

Sealed Mating Connectors: Sealed (Splash-proof) connectors were originally designed to overcome long-term integral cable sealing failures discovered in the paper industry. These connectors use "O"ring and other compression seals to prevent fluid from entering into the connector interface. Non-dielectric silicone grease is applied to both the sensor and mating connector during assembly to fill voids and prevent any vapor pressure from developing. Some designs can be used in submerged locations.

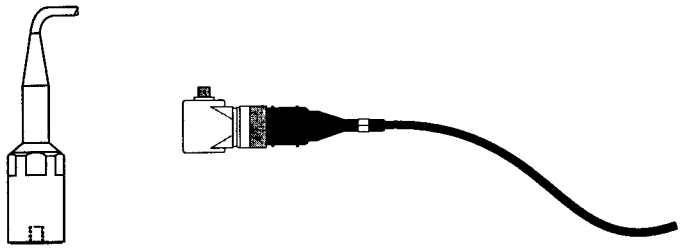


Figure 2: An older Wilcoxon 793S and a modern 797 with R6SL sealed connector

Sealed Connector and Cable Materials: The materials used on sealed connectors and cables must withstand high temperatures, humidity and a variety of process chemicals. Flouropolymer based insulator materials such as Teflon®, Tefzel®, and Viton® are recommended for these harsh industrial applications. Other cable materials such as polyurethane, polyethylene, nylons and neoprene have exhibited many problems in hot humid environments.

Flouropolymers are highly resistant to most non-flouridic chemicals and provide unparalleled temperature capabilities. Their abrasion resistance and durability are also excellent. Teflon grades are the most common can be used in applications as high as 250 °C (480 °F). Tefzel is similar in composition and exhibits excellent radiation resistance in nuclear applications. Viton is an elastomeric flouropolymer used for "O" rings and sealed cable boots.

Cable Design: Shielded-twisted-pair cables are the most common in vibration monitoring systems. They consist of a twisted-pair surrounded by a braided and/or foil shield. Signals are carried on the twisted pair and isolated from the electrical shield. The combined outer braid and conductor twist greatly attenuates electrical and magnetic interference signals. Shielded-twisted-pair cables readily mate with terminal blocks commonly found on most junction box and other signal interfaces. It is also important that stranded conductors should be used. Excessive vibration can fatigue solid core wire until failure.

Environmentally Rugged Sensor Systems: Sensor Electronics: Innovations in the design of COTS industrial accelerometer amplifiers have been driven primarily by the needs of the paper industry. Many of these developments were intended to increase reliability in this demanding environment. In general sensors designed for paper or steel roll process measurements will be more than adequate for ship-board measurements.

Long-term Temperature Survival: High temperatures can greatly reduce the life of electronic circuitry. Newer generations of amplifier designs, such as the Wilcoxon FireFET® series, are specifically designed for long-term use at high temperature. Use of such technology can be critical at some gas and steam turbine locations.

Amplifier Protection Circuitry: Survival in harsh environments requires a variety of protection circuit to be built into the sensor amplifier. Electrostatic discharge protection is the most critical near arc producing machinery and weld operations. Transient voltage suppressers prevent catastrophic or incremental amplifier damage due to most electrostatic discharge (ESD) events.

Reverse-Wire protection is recommended in multipoint installations where miswiring can be a problem. This protection circuit prevents damage if the cable leads are reversed. Sensors without protection usually fail and are non-repairable. Overcurrent protection is also available on many sensors. This circuit prevents damage if too much current is applied to the sensor. Fortunately, most vibration monitors provide standard-current-powering (SCP) and limit the current to the sensor.

Finally, overload protection is necessary to both improve signal quality and prevent amplifier failure from excessive shock. The feature is required for all accelerometers located near steam leaks, cavitating pumps and high-speed gear boxes. The noise generated by such machinery can contain tremendous amounts of high frequency energy. This energy once detected by an accelerometer can easily exceed the amplitude measurement range of the sensor. Over-excitation produces intermodulation (washover) distortion and causes false low frequency readings. Very large "ski-slopes" usually dominate the spectrum of an overloaded accelerometer. Overload protection circuits filter very high frequency signals before they can affect the amplifier. All PiezoFET® amplifiers contain this circuitry. [2]

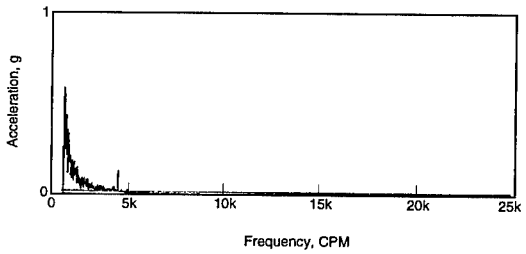


Figure 3 - Large "ski slope" response produced by a 100 mV/g accelerometer on a cavitating pump rge "ski slope" response produced by a 100 mV/g accelerometer on a cavitating pump

Environmentally Rugged Installation - Sensor Mounting: Mounting is critical for both data quality and long-term system integrity. In harsh environments finding the proper sensor location and achieving mount quality can be a difficult task. Mounting location and position of the sensor will depend on the machine construction (e.g. type of bearing) and data requirements. Installation technique can be effected by the work environment and location access.

The mounting location in harsh environments should follow normal practices common to the industry. For example, on heavily loaded bearings the sensor should be located in the load zone. However, high temperatures, chemical spray and other factors can greatly complicate the installation. Sensor mounting locations should be carefully planned before installation.

Stud mounting: Stud mounts maximize the sensor response and are the best long-term installation choice. However it is critical that proper mounting technique is used. Proper techniques maximize data quality and reduces the chance that a sensor will be knocked into the machine. Stud mounting requires that the machine location be spot-faced, drilled and tapped. Installation tools are available that can both face and drill in one operation. The use of such tools will insure that the drill hole is perfectly perpendicular to the spot face created by the end mill. The cutting tool should be made of a high speed steel for long term survivability and re-sharpening capabilities. Avoid carbide tipped tools which often crack when the rotating end mill makes contact with the machine surface

When attaching the sensor apply a thin layer of Loctite Depend® epoxy between the machine surface and the accelerometer. Loctite Depend is a "no mix adhesive" that utilizes a spray activator. Torquing the accelerometer will force the adhesive into any surface imperfections and out from around the base of the accelerometer. Acting as both a "coupling agent" and a thread lock, the adhesive improve high frequency data and mount security. [3]

Use of Magnetic and Adhesive Mounting Pads: Magnetic pads provide a location for use with flat magnets commonly used in Navy portable measurements. Adhesive pads allow sensors to be permanently mounted in locations that cannot be drilled and tapped. One side of the pad is grooved or abraded for adhesion to the machine. The other side is flat for a magnet or contains a tapped hole/integral stud for direct coupling to the sensor.

Adhesive pads are used successfully in the harshest environments. However, the surface must be properly prepared to increase adhesion. Generally this means the mounting surface must be flat, abraded, and clean of oil and debris. [4] The proper adhesive must also be selected to ensure environmental survival. Lord Chemical Versilock® 406 has proven very successful. It has been used in the paper industry for many years and shows remarkable resistance to chemical and thermal degradation. [5]

Environmentally Rugged Installation - Cable Routing: Intelligent cable installation can improve system reliability and ease operational maintenance. Cables should be routed to minimize electrical interference and reduce the possibility of physical damage. One way to decrease physical damage and improve plant safety is to leave machine entry ways uncongested. The use of low-profile sensors encourage a "cleaner" installation by allowing the cable to be routed closely along the bearing housing.

Teflon cables are very durable and can be routed with tie wraps along oil lines, railings, and other support structures. Tie wrap installations are easier to retrofit and trouble-shoot than conduit. When using tie wraps, Tefzel is recommended. Although nylon tie wraps are less expensive, they exhibit severe degradation in some applications. Finally, do not tie wrap to power lines! Route cables away from high voltage lines and motors to reduce signal interference. When necessary, cross high voltage lines at right angles.

Conduit is required in some locations, however, the added cable protection does not always justify the expense. If conduit is used it must be properly installed and maintained. Poor conduit runs can easily provide a water line directly into a termination and other electrical boxes. Cable and conduit should enter the bottom of electrical enclosures to prevent water from being piped into the box.

Conclusion: COTS technology suitable for ship-system condition assessment installations must be acquired based on many years of harsh environmental experience. Growth of maritime accelerometer installations will further highlight the importance of data quality, reliability, and long-term sensor survival.

The ship-board environmental requirements are the primary determinant for sensor selection and reliability. Use of flouropolymer cables and 316L stainless steel sensor housings are recommended to resist sea water, corrosive chemicals, lubricants and survive higher temperature machinery. Submerged and lubrication intensive sensor locations also require expert sealing technology to prevent fluid entry into the sensor. Indeed, survival in the some ship-system environments requires the same level of sealing and quality assurance techniques as the latest orbital electronics technology.

Finally, sensor location and cable routing can make the difference between a reliable program and continuing problems. Use of low profile sensors to clear machinery entry areas will reduce sensor damage during maintenance. Smart cable routing will ease maintenance and improve signal quality.

[1] "Carpenter Stainless Steels", Carpenter Technology Corporation, 1987, p.45
[2] Barrett, Richard M. "Low Frequency Machinery Monitoring: Measurement Considerations", P/PM Technology , 12/93, pp. 23-31.
[3] "Accelerometer Installations in Extreme Industrial Environments" Richard Barrett, Todd Koch, Sound & Vibration Magazine, June 1996, p.p. 20-26
[4] "Model 793 Industrial Piezoceramic Accelerometer Installation and Operation Guide" Wilcoxon Research Inc. Publication, April 1993
[[5] Versilok® modified structural adhesive, Lord Corporation, Industrial Adhesive Division, Eire, PA

OPTICAL FIBER SENSORS FOR STRAIN MEASUREMENT IN HIGH-PRESSURE AIR COMPRESSOR BOLTS

Vikram Bhatia, Marten J. de Vries, Kent A. Murphy, and Richard O. Claus

Fiber & Electro-Optics Research Center, Bradley Department of Electrical Engineering
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0358

Christopher P. Nemarich

Naval Surface Warfare Center, Carderock Division, Code 8530
3A Leggett Circle, Annapolis, MD 21402-5067

Abstract: Optical fiber sensors have found numerous applications in industrial, military and commercial sectors in the last two decades. We discuss the principle of operation of two basic types of fiber optic sensors for on-line strain measurements. Interferometric and grating-based sensors are evaluated on the basis of simplicity of fabrication, signal processing techniques, strain resolution and cross-sensitivity to temperature fluctuations. Extrinsic and intrinsic Fabry-Perot interferometers and Bragg and long-period gratings are investigated to determine their feasibility for potential strain measurement in high-pressure air compressor bolts.

Key Words: Gratings; Interferometers; Optical fiber sensors; Strain measurements.

Introduction: Strain measurements using optical fiber sensors in both embedded and surface-mounted configurations have been reported by researchers in the past [1,2,3]. Fiber optic sensors are small in size, immune to electromagnetic interference and can be easily integrated with existing optical fiber communication links. Such sensors can typically be easily multiplexed, resulting in distributed networks that can be used for health monitoring of integrated, high-performance materials and structures. Optical fiber sensors for strain measurements should possess certain important characteristics. These sensors should either be insensitive to ambient fluctuations in temperature and pressure, or should have demodulation techniques that compensate for changes in the output signal due to the undesired perturbations. The sensor signal should itself be simple and easy to demodulate. Nonlinearities in the output demand expensive decoding procedures or require pre-calibrating the sensor. The sensor should ideally provide an absolute and real-time strain measurement in a form that can be easily processed. For environments where large strain magnitudes are expected, the sensor should have a large dynamic range while at the same time maintaining the desired sensitivity.

We discuss the operation of interferometric and grating-based optical fiber sensors for strain measurements. Extrinsic and intrinsic Fabry-Perot interferometers and Bragg and long-period gratings are investigated to determine their feasibility for potential strain measurement in high-pressure air compressor bolts. Each of the four sensor types is

first detailed individually. Particular attention is paid to the operating principle, sensor fabrication, strain sensitivity and dynamic range, cross-sensitivity to temperature and transverse strain components, and signal demodulation techniques. This performance comparison is then used to down-select sensors for final tests.

Extrinsic Fabry-Perot Interferometric Sensor: The extrinsic Fabry-Perot interferometric (EFPI) sensor, proposed by Murphy *et al.*, is one of the most popular fiber optic sensors used for applications in health monitoring of smart materials and structures [3]. As the name suggests the EFPI is an interferometric sensor in which the detected intensity is modulated by the parameter under measurement. The simplest configuration of an EFPI is shown in Fig. 1.

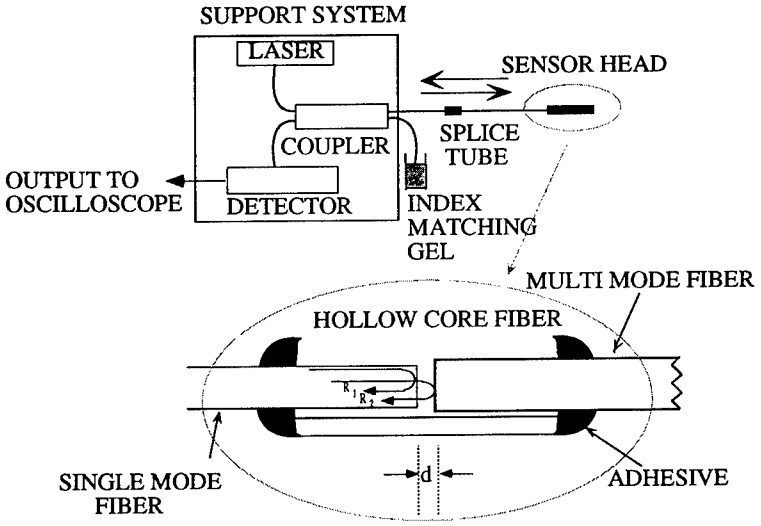


Fig. 1. The extrinsic Fabry-Perot interferometric (EFPI) sensing system.

The EFPI system consists of a single mode laser diode that illuminates a Fabry-Perot cavity through a fused biconical tapered coupler. The cavity is formed between an input single mode fiber and a reflecting single mode or multimode fiber. Since the cavity is external to the lead-in/lead-out fiber the EFPI sensor is independent of ambient temperature fluctuations and transverse strain. The input fiber and the reflecting fiber are aligned by using a hollow core silica fiber. For uncoated fiber ends, a 4% Fresnel reflection results at both ends of the cavity. The first reflection, R_1, called the reference reflection, is independent of the applied perturbation. The second reflection, R_2, termed the sensing reflection, is dependent on the length of the cavity, d, which in turn is modulated by the applied perturbation. These two reflections interfere (provided $2d<L_C$, the laser diode's coherence length) and the intensity I at the detector varies as a function of the cavity length,

$$I = I_0 \cos\left(\frac{4\pi}{\lambda} d\right), \qquad (1)$$

where, I_0 is the maximum value of the output intensity and λ is the laser diode center wavelength.

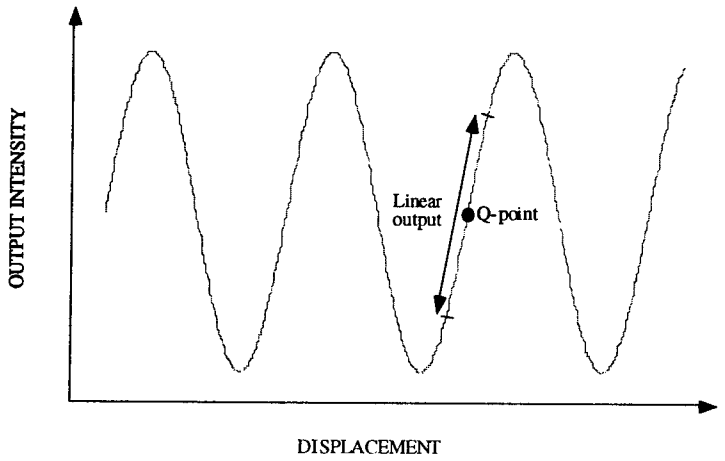


Fig. 2. The EFPI transfer function curve.

The typical EFPI transfer function curve is shown in Fig. 2. Small perturbations that result in operation around the quiescent-point or Q-point of the sensor lead to a linear variation in output intensity. A fringe in the output signal is defined as the change in intensity from a maximum to a maximum or from a minimum to a minimum. Each fringe corresponds to a change in the cavity length by one half of the operating wavelength, λ. The change in the cavity length, Δd, is then employed to calculate the strain using the expression,

$$\varepsilon = \frac{\Delta d}{L}, \tag{2}$$

where, L is defined as the gage-length of the sensor and is typically the distance between two points where the input and reflecting fibers are bonded to the hollow-core fiber.

The EFPI sensor has been extensively used for measuring fatigue loading on F-15 aircraft wings, detection of crack formation and propagation in civil structures and cure and lifetime monitoring in concrete and composite specimens [1,2]. The temperature-insensitivity of this sensor makes it attractive for a large number of applications. The EFPI sensor is capable of measuring sub-angstrom displacements with strain resolution better than 1 $\mu\varepsilon$ and a dynamic range greater than 10,000 $\mu\varepsilon$. Although the change in output intensity of the EFPI is non-linear corresponding to the magnitude of the parameter being measured, for small perturbations its operation can be limited to that around the Q-point of the transfer function curve. Moreover, the large bandwidth available with this sensor simplifies the measurement of highly cyclical strain. The EFPI sensor is capable of providing single-ended operation and is hence suitable for applications where access to the test area is limited. The sensor requires simple and inexpensive fabrication equipment and an assembly time of less than 10 minutes. Additionally, since the cavity is external to the fibers, transverse strain components that tend to influence intrinsic sensors through the Poisson's effect have negligible effect on the EFPI sensor output [4]. The sensitivity to only axial strain and insensitivity to input polarization state have made the EFPI sensor the most preferred fiber optic sensor for

embedded applications. Thus overall the EFPI sensing system is very well suited to measurement of small magnitudes of cyclical strain.

Intrinsic Fabry-Perot Interferometric Sensor: The intrinsic Fabry-Perot interferometric (IFPI) sensor is similar in operation to its extrinsic counterpart but significant differences exist in configurations of the two sensors [5]. A laser diode is used as the optical source to one of the input arms of a bi-directional 2x2 coupler. The Fabry-Perot cavity is formed by fusing a small length of a single mode fiber to one of the output legs of the coupler. The reference and sensing reflections interfere at the detector face to provide a sinusoidal intensity variation. The cavity can also be obtained by introducing two Fresnel reflectors along the length of a single fiber. Since the cavity is formed within an optical fiber, changes in the refractive index of the fiber due to the applied perturbation can significantly alter the phase of the sensing signal. Thus the intrinsic cavity results in the sensor being sensitive to ambient temperature fluctuations and all six states of strain.

The IFPI sensor, like all other interferometric signals has a non-linear output that complicates the measurement of large magnitude strain. This can again be overcome by operating the sensor in the linear regime around the Q-point of the sinusoidal transfer function curve. The main limitation of the IFPI strain sensor is that the photoelastic effect-induced change in index of refraction results in a non-linear relationship between the applied perturbation and the change in cavity length. In fact for most IFPI sensors the change in propagation constant of the fundamental mode dominates the change in cavity length. Thus IFPIs are highly susceptible to temperature changes and transverse strain components [4]. In embedded applications the sensitivity to all six strain state can result in erroneous outputs. The fabrication process of an IFPI strain sensor is more complicated than that of the EFPI sensor since the sensing cavity has to be formed within the optical fiber by some special procedure. The strain resolution of the IFPIs is also expected to be around 1 $\mu\varepsilon$ with an operating range greater than 10,000 $\mu\varepsilon$. IFPI sensor also suffer from drift in the output signal due to variations in the polarization state of the input light.

Thus the preliminary analysis shows that the extrinsic version of the Fabry-Perot optical fiber sensor seems to have an overall advantage over its intrinsic version. The extrinsic sensor has negligible cross-sensitivity to temperature and transverse strain. Although the strain sensitivity, dynamic range and bandwidth of the two sensors are comparable, the IFPIs can be expensive and cumbersome to fabricate due to the intrinsic nature of the sensing cavity.

The extrinsic and intrinsic Fabry-Perot interferometric sensors possess non-linear sinusoidal outputs that complicate the signal processing at the detection end. Although intensity-based sensors have a simple output variation, they suffer from limited sensitivity to strain or other perturbations of interest. Grating-based sensors [6] have recently become popular as transducers that provide wavelength-encoded output signals that can typically be easily demodulated to derive information about the perturbation under investigation. We first discuss the basic operating mechanism, advantages and drawbacks of the Bragg grating sensing technology. These sensors are then compared to the recently developed long-period gratings [7] in terms of fabrication process, cross-sensitivity to other parameters and simplicity of signal demodulation.

Fiber Bragg Grating Sensor: Fiber Bragg gratings are based on the photosensitivity of germanosilicate fiber core to intense ultra-violet light. A refractive index modulation of the core causes light at a particular wavelength to be reflected back. Bragg gratings

are based on the phase-matching condition between spatial modes propagating in optical fibers,

$$\Delta\beta = \frac{2\pi}{\Lambda}, \tag{3}$$

where, $\Delta\beta$ is the difference in the propagation constants of the two modes involved in mode coupling (both assumed to travel in the same direction).

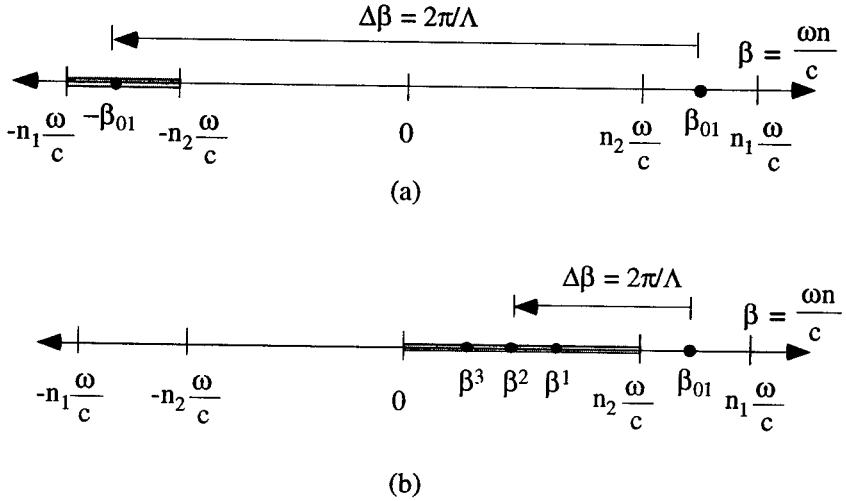


Fig. 3. Depiction of mode coupling in (a) Bragg gratings and, (b) long-period gratings. The differential propagation constant, $\Delta\beta$, determines the grating periodicity.

Fiber Bragg gratings (FBGs) involve the coupling of the forward-propagating LP_{01} mode to the reverse-propagating LP_{01} mode [8]. To satisfy the phase-matching condition,

$$\lambda_B = 2\Lambda n_{eff}, \tag{4}$$

where, λ_B is termed the Bragg wavelength and n_{eff} is the effective index of the fundamental mode. Bragg wavelength is the wavelength at which the forward-propagating LP_{01} mode couples to the reverse-propagating LP_{01} mode. This coupling is wavelength dependent since the propagation constants of the two modes are a function of the wavelength. Hence, if a FBG is interrogated by a broadband optical source, the wavelength at which phase-matching occurs is reflected back. This wavelength is a function of the grating periodicity (Λ) and the effective index (n_{eff}) of the fundamental mode (Eq. (4)). Since strain and temperature effects can modulate both these parameters, the Bragg wavelength shifts with these external perturbations. This spectral shift is utilized to fabricate FBGs for sensing applications.

Fig. 3 (a) shows the mode coupling mechanism in fiber Bragg gratings using the β-plot. Since the difference in propagation constants ($\Delta\beta$) between the modes involved in coupling is large, we see from Eq. (4) that only a small value of periodicity, Λ, is needed to induce this mode coupling. Typically for telecommunication applications the value of λ_B is around 1.5 μm. From Eq. (4), Λ is determined to be 0.5 μm (for n_{eff}=1.5).

Due to the small periodicities (of the order of 1 μm) FBGs are classified as short-period gratings.

From Eq. (4) we see that a change in the value of n_{eff} and/or Λ can cause the Bragg wavelength, λ, to shift. This fractional change in the resonance wavelength, $\Delta\lambda/\lambda$, is given by the expression,

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta\Lambda}{\Lambda} + \frac{\Delta n_{eff}}{n_{eff}}, \tag{5}$$

where, $\Delta\Lambda/\Lambda$ and $\Delta n_{eff}/n_{eff}$ are the fractional changes in the periodicity and the effective index, respectively. The relative magnitudes of the two changes depend on the type of perturbation the grating is subjected to, and for most applications the effect due to change in effective index is the dominating mechanism.

Any axial strain, ε, applied to the grating changes the periodicity and the effective index and results in a shift in the Bragg wavelength, given by,

$$\frac{1}{\lambda}\frac{\Delta\lambda}{\varepsilon} = \frac{1}{\Lambda}\frac{\Delta\Lambda}{\varepsilon} + \frac{1}{n_{eff}}\frac{\Delta n_{eff}}{\varepsilon}. \tag{6}$$

The first term on the right-hand side is unity, while the second term has its origin in the photoelastic effect. An axial strain on the fiber serves to change the refractive index of both the core and the cladding. At 1550 and 1300 nm the shifts in the resonance wavelength are 11 nm/%ε and 9 nm/%ε, respectively. With temperature a FBG at 1500 nm shifts by 1.3 nm for every 100 °C rise in temperature.

A major limitation of Bragg grating sensors is the complex and expensive fabrication technique. Although side-writing is commonly being used to manufacture these gratings, the requirement of expensive phase masks increases the cost of the sensing system. In the interferometric technique, stability of the set up is a critical factor to obtain high quality gratings.

The second major drawback of Bragg gratings is their limited spectral shift with strain and temperature. The limited wavelength shift demands high-resolution spectrum analyzers to monitor the grating spectrum. Kersey *et al.* have proposed an unbalanced Mach-Zender interferometer to detect the perturbation-induced wavelength shift [9]. Two unequal arms of the Mach-Zender interferometer are excited by the backreflection from a Bragg grating sensor element. Any change in the input optical wavelength modulates the phase difference between the two arms and results in a time-varying sinusoidal intensity at the output. This interference signal can be related to the shift in the Bragg peak and hence, the magnitude of the perturbation can be obtained. The unbalanced interferometers are also susceptible to external perturbations and hence need to be isolated from the parameter under investigation. Moreover, the non-linear output might require fringe counting equipment which can be complex and expensive. Additionally, a change in the perturbation polarity at the maxima or minima of the transfer function curve will not be detected by this demodulation scheme. To overcome this limitation two unbalanced interferometers might have to be employed for dynamic measurements.

The cross-sensitivity to temperature fluctuations leads to erroneous strain measurements in applications where the ambient temperature has a temporal variation. So a reference grating that measures the temperature change has to be utilized to compensate for the output of the strain sensor. Recently, temperature-independent

sensing has been demonstrated using chirped gratings written in tapered optical fibers [10].

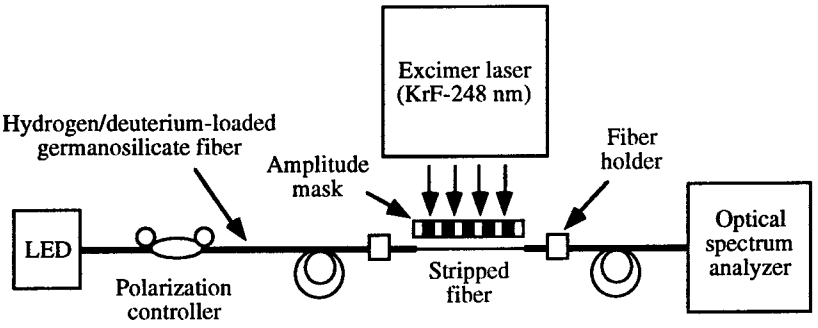


Fig. 4. Set up to fabricate long-period gratings using an amplitude mask.

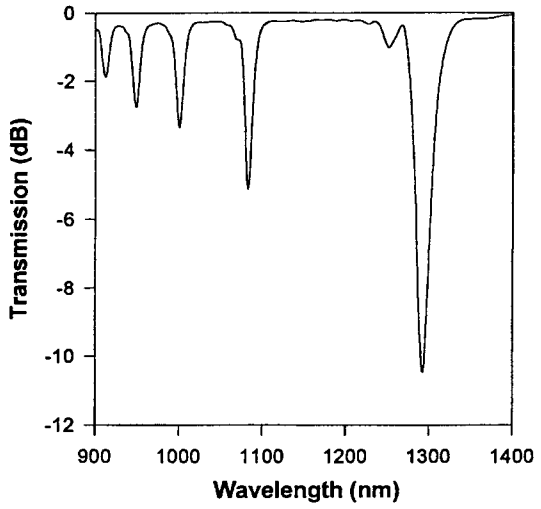


Fig. 5. Transmission spectrum of a long-period grating written in Corning FLEXCOR fiber with period, Λ=198 μm. The discrete, spiky loss bands correspond to the coupling of the fundamental guided mode to discrete cladding modes.

Long-Period Grating Sensor: For conventional fiber Bragg gratings the coupling of the forward propagating LP_{01} mode occurs to the reverse propagating LP_{01} mode ($\beta=-\beta_{01}$). Since, $\Delta\beta$ is large in this case (Figure 3 (b)), the grating periodicity is small, typically of the order of one micrometer. Unblazed long-period gratings couple the fundamental mode to the discrete and circularly-symmetric, forward-propagating cladding modes ($\beta=\beta^n$), resulting in smaller values of $\Delta\beta$ (Figure 3 (b)) and hence periodicities ranging in hundreds of micrometers [10]. The cladding modes attenuate rapidly as they propagate along the length of the fiber due to the lossy cladding-coating interface and bends in the fiber. Since $\Delta\beta$ is discrete and a function of the wavelength,

this coupling to the cladding modes is highly selective, leading to a wavelength-dependent loss. As a result, any modulation of the core and cladding guiding properties modifies the spectral response of long-period gratings, and this phenomenon can be utilized for sensing purposes. Moreover, since the cladding modes interact with the fiber jacket or any other material surrounding the cladding, changes in the properties of these ambient materials can also be detected.

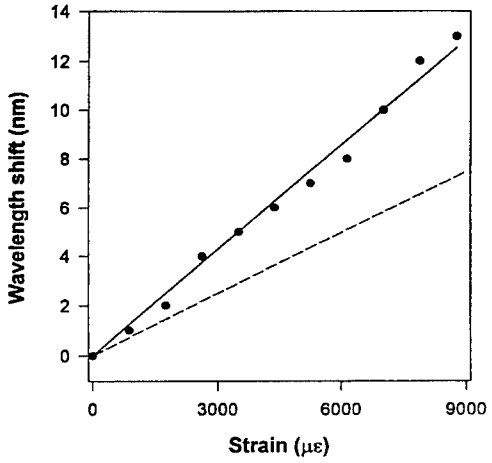


Fig. 6. Shift in the highest order resonance band with strain for a long-period grating written in fiber D (circles). Also depicted is the shift for a conventional Bragg grating (dashed line).

To fabricate long-period gratings hydrogen-loaded (3.4 mole%), germanosilicate fibers are exposed to 248 nm UV radiation form a KrF excimer laser, through a chrome-plated amplitude mask possessing a periodic rectangular transmittance function. Fig. 4 shows the set up used to fabricate the gratings. The laser was pulsed at 20 Hz with a 8 ns pulse duration. The typical writing times for an energy of 100 mJ/cm^2/pulse and a 2.5 cm exposed length vary between six to fifteen minutes for different fibers. Fig. 5 depicts the typical transmittance of a grating. Various attenuation bands correspond to coupling to discrete cladding modes of different orders. A number of gratings can be fabricated at the same time by placing more than one fiber behind the amplitude mask. Moreover, the stability requirements during the writing process are not so severe as those for short-period Bragg gratings.

For coupling to the highest-order cladding-mode, the maximum isolation (loss in transmission intensity) is typically in the 5 to 20 dB range on wavelengths depending on fiber parameters, duration of UV exposure and mask periodicity. The desired fundamental coupling wavelength can easily be varied by using an inexpensive amplitude masks of different periodicities. The insertion loss, polarization-mode dispersion, backreflection, and polarization-dependent loss of a typical grating are 0.2 dB, 0.01 ps, -80 dB, and 0.02 dB, respectively [7]. The negligible polarization sensitivity and backreflection of these devices eliminates the need for expensive polarizers and isolators.

The strain sensitivity of a grating written in Flexcor fiber fibers was determined by axially straining the gratings between two longitudinally separated translation stages. The shift in the peak loss wavelength as a function of the applied strain is depicted in Fig. 6 along with that for a Bragg grating (about 8.3 nm/%ε at 1250 nm). Fiber D has β=15.2 nm/%ε which gives it a strain coefficient that is slightly less than twice that of a Bragg grating at the same wavelength. The strain resolution of this fiber for a 0.1 nm detectable wavelength shift is 65.75 με. The temperature coefficient for a grating written in Flexcor fiber is 0.111 nm/°C which is almost an order of magnitude higher than that of a Bragg grating. This implies that a larger cross-sensitivity to temperature is expected for long-period gratings while measuring strain. Also long-period gratings are extremely sensitive to bends in the fiber and hence the device has to be kept straight during the measurement process. The effect of transverse strain components on the grating spectrum is presently under investigation.

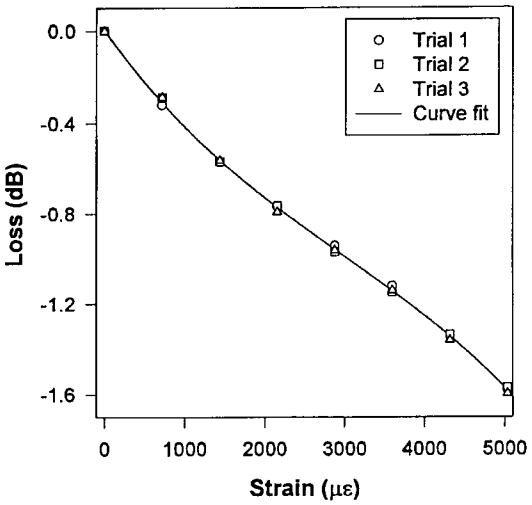


Fig. 7 The change in the grating transmission at 1317 nm as a function of strain for three different trials.

The demodulation scheme of a sensor determines the overall simplicity and sensitivity of the sensing system. Short-period Bragg grating sensors were shown to possess signal processing techniques that are complex and expensive to implement. We now present a simple demodulation method to extract information from long-period gratings. The wide bandwidth of the resonance bands enables the wavelength shift due to the external perturbation to be converted into an intensity variation that can be easily detected. For a grating with λ_p=1306 nm, a laser diode centered at 1317 nm was used as optical source and the change in transmitted intensity was monitored as a function of applied strain. The transmitted intensity is plotted in Fig. 7 for three different trials. The repeatability of the experiment demonstrates the feasibility of using this simple scheme to utilize the high sensitivity of long-period gratings.

Conclusions: We have compared the performance of four different fiber optic sensors for strain measurements. While the two interferometric sensors possess high resolution, their non-linear output require complex signal processing. The extrinsic version of the

fiber Fabry-Perot interferometric sensor has advantages over its intrinsic counterpart due to smaller cross-sensitivity to temperature and transverse strain. The grating-based sensors offer multiplexing capabilities which brings down the cost per strain channel. Long-period grating sensors offer simpler demodulation methods and inexpensive fabrication. Bragg grating strain sensors provide smaller cross-sensitivity to temperature and bends during strain measurement.

Based on the above comparison EFPI and Bragg grating sensors promise the most favorable performance for strain measurements in environments with temperature fluctuations. Future work would focus on developing attachment techniques to mount these sensors on air compressor components and testing them under cyclical loading on mechanical load frames.

References

1. K. A. Murphy, M. F. Gunther, A. M. Vengsarkar, and R. O. Claus, "Fabry-Perot fiber optic sensors in full-scale fatigue testing on an F-15 aircraft," *Applied Optics,* vol. 31, pp. 431-433, 1991.

2. V. Bhatia, C. A. Schmid, K. A. Murphy, R. O. Claus, T. A. Tran, J. A. Greene, and M. S. Miller, "Optical fiber sensing technique for edge-induced and internal delamination detection in composites," *Journal of Smart Materials and Structures*, vol. 4, 1995.

3. R. O. Claus, M. F. Gunther, A. Wang, and K. A. Murphy , "Extrinsic Fabry-Perot sensor for strain and crack opening displacement measurements from -200 to 900 °C," *Smart Materials and Structures*, vol. 1, pp. 237-242, 1992.

4. J. Sirkis, "Phase-strain-temperature model for structurally embedded interferometric optical fiber strain sensors with applications," *Fiber Optic Smart Structures and Skins IV*, SPIE, vol. 1588.

5. C. E. Lee and H. F. Taylor, "Fiber-optic Fabry-Perot temperature sensor using a low-coherence light source," *Journal of Lightwave Technology*, vol. 9, pp. 129-134, 1991.

6. K. O. Hill, Y. Fuijii, D. C. Johnson, and B. S. Kawasaki, "Photosensitivity in optical fiber waveguides: Applications to reflection filter fabrication," *Applied Physics Letters*, vol. 32, p. 647, 1978.

7. A. M. Vengsarkar, P. J. Lemaire, J. B. Judkins, V. Bhatia, J. E. Sipe, and T. E. Ergodan, "Long-period fiber gratings as band-rejection filters," *Journal of Lightwave Technology*, vol. 14, 58, 1996.

8. G. Meltz, W. W. Morey and W. H. Glenn, "Formation of Bragg gratings in optical fibers by transverse holographic method," *Optics Letters*, vol. 14, p. 823, 1989.

9. A. D. Kersey and T. A. Berkoff, "Fiber-optic Bragg-grating differential-temperature sensor," *IEEE Photonics Technology Letters*, vol. 4, pp. 1183-1185, 1992.

10. M. G. Xu, L .Dong, L. Reekie, J. A. Tucknott and J. L. Cruz, "Chirped fiber gratings for temperature-independent strain sensing," in *Proceedings of the First OSA Topical Meeting on Photosensitivity and Quadratic Nonlinearity in Glass Waveguides: Fundamentals and Applications*, paper PMB2, 1995.

REAL TIME INTERFEROMETRIC ANALYSIS OF STRAINED SILICONE FILM

G. E. Dovgalenko, Y. I. Onischenko, A. V. Kniazkov, M. S. Haque, H. A. Naseem, G. J. Salamo
University of Arkansas, Fayetteville, AR 72701, USA

Abstract: Time-dependent stress change in thin SiO_2 films on Si substrate was investigated after deposition using high stability portable holographic interferometer. In parallel, measurement was done in a Tencor stress measurement system. A stress relaxation phenomenon in this film thickness of 0.5 μm on Si wafer has been observed. The spiral anisotropy of w displacement component have been discovered. Model for explanation of this phenomena has been proposed. The advantages of the proposed measured technique and results are discussed.

Key Words: anisotropy, films, interferometry, portability, stress relaxation.

INTRODUCTION: The measurement of stress in thin films is crucial to the successful fabrication of microelectronic elements, masks for X-ray lithography, etc. Distortion due to stress can render a mask unusable if the stress is not minimized. The local stress in thin film for example SiO_2 on Si substrate can made damage in electronic microchips. The ability to accurately measure stress is required of course, before it can be controlled. Thin films deposited on hot substrates get a resultant stress in the film because of the difference in the thermal expansion coefficients of the film and the substrate. The deposited films will have another component of stress depending on the film microstructure and type and concentration of incorporated impurities. Traditionally bending plate deflection technique is used for thin film stress measurement where change in bow of the substrate due to deposited film is measured. Tencor FLX2320 system uses dual wavelength laser deflection technique. Unfortunately this device provides measurements only for good reflecting surfaces. For diffusely reflected surfaces such as SiO_2 the amount of reflected light is not enough for stress measurements. It provides only average stress measurement. This excludes information concerning of local stress distribution on the surface which is important for optimization of deposition condition and quality test. This device needs recalibration in time, is expensive, not portable, operates in quasi-real time.
The challenge of thin film fabrication technology is stress measurement [1]using a portable, economical real-time technique which allows measurements of local stress on 2-D map surface. We demonstrate a prototype of interferometer which satisfies condition listed above.

EXPERIMENT A prototype of the device for stress testing of thin film is presented on Fig.1. Using a single-beam construction of interferometer and a reversible thermoplastic camera which is fixed on a rigid frame, we can eliminate a vibroprotection system [2].Under this condition it is possible to create high stable real-time interferometer for

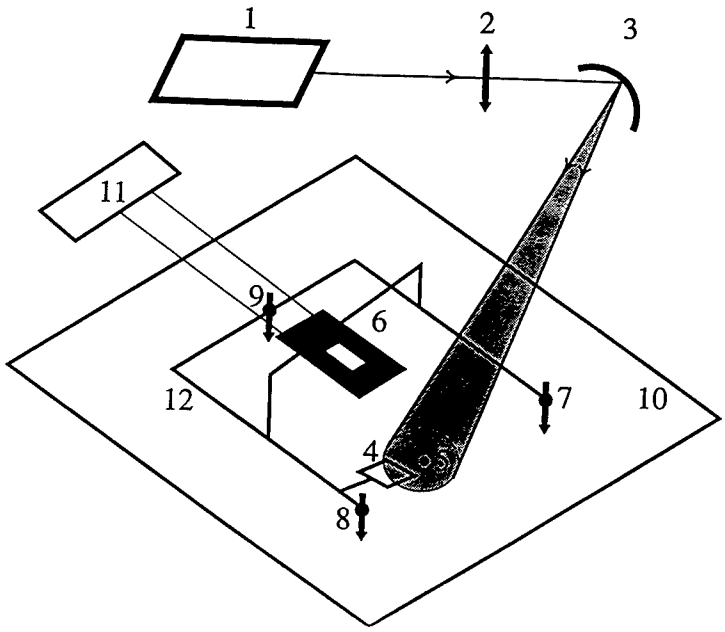


Fig. 1. Prototype device for real time stress testing of thin films.
1- laser (He-Ne), 2- lens, 3 - spherical mirror, 4 - mirror, 5 - testing zone,
6 - thermoplastic camera, 7, 8, 9 - adjustable supports, 10 - table surface,
11- high voltage supply, 12- rigid frame

testing long time stress relaxation processes and displacement component of surface which depends on stress:

$$\sigma = 2\,We h^2/(1-\nu)6a^2t \qquad (1)$$

where E is Young's module, ν is Poisson's ratio., W is substrate displacement, t- is the film thickness, h-is substrate thickness. The diffracted efficiency of thermoplastic film is 17 % using 20 mW CW He-Ne laser. The construction on Fig.1 is extremely time stable and allows to provide long time dependent measurements with accuracy of order of $\lambda/2$, where λ is laser wavelength. The surface movements of the displacement vector **L**(U,V,W) in the direction **R**(X,Y,Z) are related to the phase change ϕ as [3]:

$$\phi = n\pi = K_xU + K_yV + K_zW \qquad (2)$$

where n is fringe number of interferogram, U,V,W - are Cartesian components of the displacement vector, K_x, K_y, K_z - are Cartesian components of sensitivity vector **K.** With this notation, dark fringes are produced for n = 1, 3, 5, ... and light fringes for n = 0,2,4... One can construct interferometric set up sensitive only to W component displacement. In this case K = (0,0,K_z). Using (1) and (2) we can compute stress σ. In our experiments parameter h was 360 μm. Using Tencor FLX2320 system we can calibrate the initial stress as a reference value at our experiments. One fringe is equal to 4.5 MPa. Fig. 2 and Fig.3 present time-dependent stress measurements. We discovered the relaxation process in time for 0.5 μm SiO_2 film deposited on Si wafer. For wafer with bad flat surface (flat deviation of order of 5 μm) spiral anisotropy displacement (SAD) W_{max} has been discovered (Fig.4.) For wafer with good flat surface (flat deviation of order of 0.8 μm) the spiral anisotropy displacement W_{max} did not observed.

MODEL: For explanation of this phenomenon we suggested that SAD depends on local displacement caused by nonuniform distribution of stress on the film surface.
Large nonuniform stress causes large SAD. Our method allows to visualize 2-D zones of SAD on the surface film and evaluate W_{max}. On Fig.6 one can see a large nonsymmetry of W_{max} for wafer with big flat deviation and small one of W_{max} Fig. 5 for wafer with small flat deviation. This test was performed on cut wafers after deposition.

CONCLUSION: A portable, cost effective system for visualization of local non-uniform stress of diffusely reflective surface has been created. The phenomenon of stress relaxation in 0.5 μm thick film of SiO_2 deposited on 360μm Si wafer has been observed. The spiral anisotropy displacement of W_{max} has been discovered. The model for explanation of this phenomenon is proposed. Application of this result for quality test of deposited thin films is presented.

REFERENCES:

1. M. Kujawinska, L. Salbut, "New trends in instrumentation and application of automated grating interferometry", Proc. SPIE, vol.2545, pp.1-16, July 1995, USA.

2. G. Dovgalenko et al., "Portable holographic interferometric device for aircraft inspection", Proc. SPIE, Vo..2545, pp.265-275, July 1995, USA.
3. A.Makino and D. Nelson, "Residual-stress Determination by Single-axis Holographic Interferometry and Hole Drilling - Part I: Theory", Experimental Mechanics, Vol.34, # 1, pp.66-78, 1994, USA.

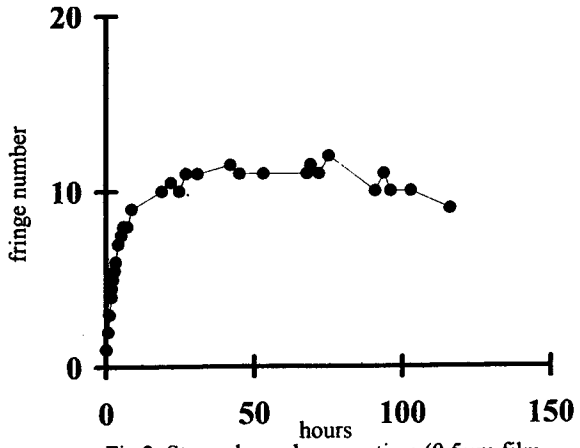


Fig.2. Stress dependence vs time (0.5μm film thickness)

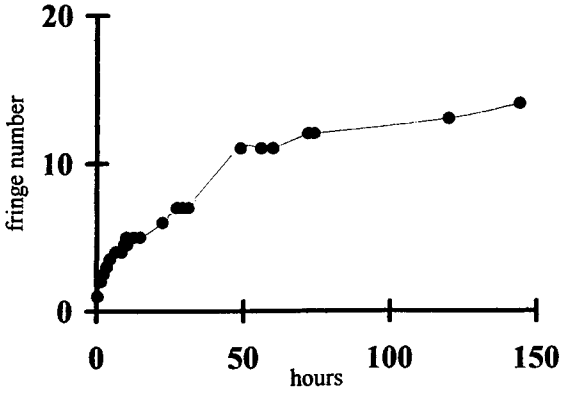


Fig. 3. Stress dependence vs time (1 μm film thickness)

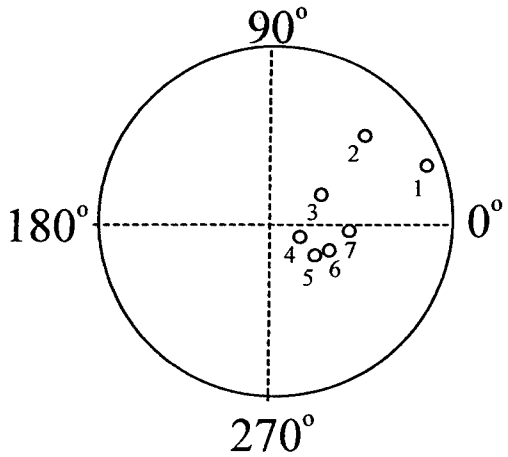


Fig. 4. Spiral anisotropy of maximum w displacement in 0.5 μm thick film in polar axes. 1- 15 minutes after deposition, 7 - two weeks after deposition.

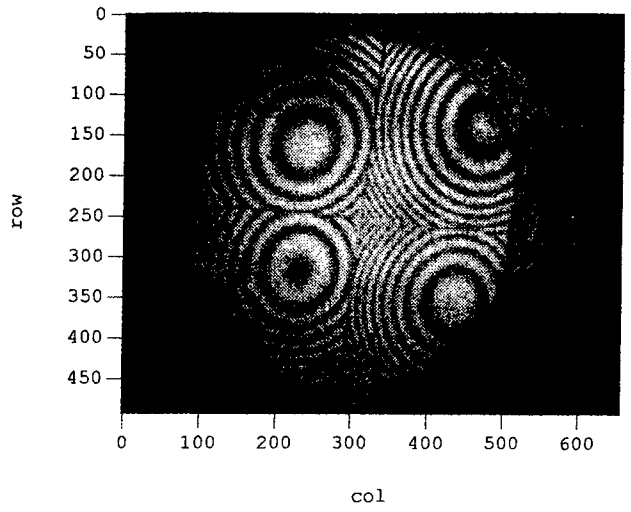


Fig.5. Four parallel independent interferogram of 0.5 μm SiO_2 film on Si surface with flat deviation of order of ± 0.8μm.

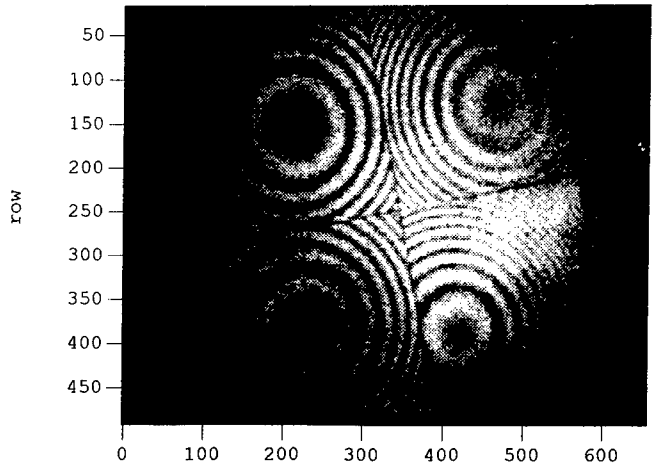


Fig. 6. Four parallel independent interferogram of 0.5 μm SiO_2 film on Si surface with flat deviation of order of ± 5μm.

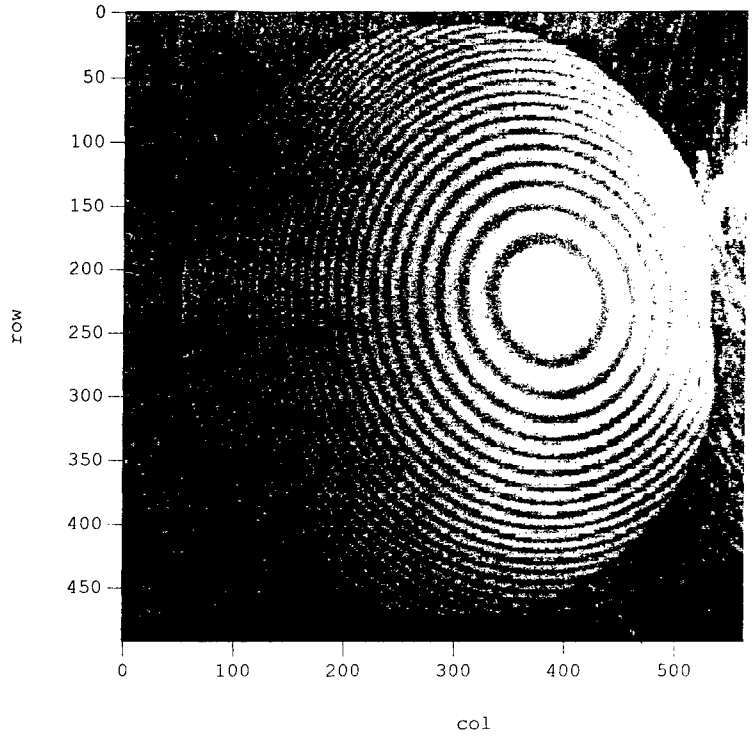


Fig.7 The interferogram of final phase of spiral anisotropy of W_{max} in 0.5μm SiO_2 film with large flat deviation ± 5 μm.

MFPT

DIAGNOSTICS AND PROGNOSTICS II

CoChairs: **Mark L. Hollins**
Naval Air Warfare Center

Stan L. Bunch
Lockheed Martin Energy Systems

AR MODELING WITH DIMENSION REDUCTION FOR MACHINERY FAULT CLASSIFICATION

Amulya K. Garga
Bryan T. Elverson
Derek C. Lang

Applied Research Laboratory
The Pennsylvania State University
P.O. Box 30
State College, PA 16804-0030
Phone: +1 814-863-5841
Fax: +1 814-863-0673
Email: garga@psu.edu

Abstract: Autoregressive (AR) and autoregressive moving-average (ARMA) modeling have been applied to the modeling of machinery signals in the past. However, to adequately model the signals large model orders have been used due to the presence of intermodulation and other sources of noise. We have developed a new approach for machinery fault classification through AR modeling followed by dimension reduction, utilizing techniques prevalent in the statistics and pattern recognition communities. The goal is to enhance those features of the AR-coefficient vector that contribute the most to the classification of faults. We describe the details of our hybrid approach and illustrate it by analyzing the Westland helicopter transmission vibration signal data set. The data consist of 8 accelerometer signals for each of 68 no-fault and seeded-fault runs on the aft transmission of a CH-46E helicopter. We first show that each data channel exhibits stationarity, which permits us to split each file into multiple 1-second segments for processing. This reduces processing time and provides for additional training data, allowing us to obtain statistically significant results. Classification performance is very dependent on the accelerometer used for generating features; as expected, the accelerometer closest to the faulty component provides the most information about the fault. We then compare the effect of different dimension-reduction techniques on fault classification. We show that for higher-order AR vectors all the dimension reduction techniques give comparable results. But, for lower-order AR vectors non-parametric methods result in greater fault separability at the cost of greater computation. Finally, we suggest a fault classification method suitable for real-time operation where low order AR modeling is followed by dimension reduction and non-parametric classification. Here, each part is a linear operation, and therefore, can be combined into a single linear operation, thereby saving greatly in complexity. These results should further the performance of CBM systems and bring us closer to adequately addressing prognostics for complex machinery.

Key Words: Autoregressive modeling; condition-based maintenance; dimension reduction; fault classification; neural networks

Introduction:

Condition-based maintenance (CBM) of mechanical systems requires the knowledge of impending fault conditions some time before they become catastrophic. Then maintenance actions can be performed to remove the faulty components. The goal of a CBM system is to *monitor* the operation of a complex mechanical equipment and provide the operator with an *accurate assessment* of the system's current health (state), and, if possible a *prediction* of the remaining useful life of the equipment [5, 7]. Successful CBM systems would employ improved robust sensors, real-time techniques for processing sensor data, fusion techniques for combining resulting refined information and micro and macro level models, and automated reasoning techniques to interpret the results of the previous analysis in the context of the operational environment and the task goals. CBM represents a migration from the traditional maintenance approaches where maintenance is either *restorative*, i.e. performed after a failure occurs, or *schedule-based*, i.e. performed after pre-determined intervals, which are typically based on worst-case mortality statistics [7]. Though the traditional approaches usually avoid catastrophic failure, they lead to unnecessary downtime, and either unacceptable risk (with restorative maintenance) or inefficient utilization of resources (with schedule-based maintenance).

Crucial to CBM is the ability to monitor the machine's condition and accurately assess fault conditions. This requires monitoring of the machine's vibration, acoustic emission, temperature, and pressure, etc. However, to prevent information overload it is important that the sensor data be presented in a more refined and concise manner. This can be accomplished by *sensor data fusion* to remove redundant information and to combine imprecise information, in order to obtain accurate and concise information. This is possible because different sensors provide information about different aspects of a complex system and data fusion utilizes the most reliable information from each sensor to provide more accurate information about the complex system than is possible to obtain with any one sensor [6]. Of course, the better the indications of fault conditions the more successful would the results of maintenance be. The are of automated reasoning for CBM has also been receiving increasing attention recently [3-5].

At the Pennsylvania State University a number of experiments are being conducted to develop techniques to diagnose and predict fault conditions for mechanical systems [5, 7]. In this paper we show how the use of autoregressive modeling followed by dimension reduction results in improved fault classification. In the next section, we provide some background on autoregressive modeling of sensor signals. Then, the new approach for fault classification is presented and illustrated with the use of Westland Helicopters CH-46E helicopter aft transmission vibration data. (The data can be explored over the web at *http://wisdom.arl.psu.edu*) We conclude with a summary and brief description of further research directions.

Background:

Autoregressive (AR) and autoregressive moving-average (ARMA) modeling have been applied to the modeling of machinery signals and other almost periodic signals with reasonable success. However, to adequately model the signals large model orders have been used due to the presence of intermodulation and other sources of noise [1]. The goal is to model a time-series signal *x(t)* such as one shown in Figure 1 using an autoregressive model:

$$y(t) = a_1 y(t-1) + a_2 y(t-1) + \cdots + a_n y(t-1)$$

so that the error *e(t)* = *x(t)* - *y(t)* is minimized according to some performance metric; a commonly used metric is the minimum mean square of the difference. For details on how to obtain the AR coefficients see [8, Ch. 1-2] or a book on time-series modeling. The main difference among the various techniques is due to the way the original and the modeled time-series are compared. One way to consider the above equation is as a one-step ahead predictor (in that case the past samples of the *y* are replaced with the samples of x), and then the minimization criterion is the prediction error.

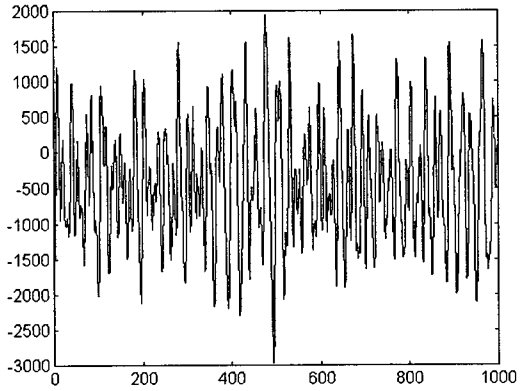


Figure 1. Example of a time-series signal to be modeled.

When using this approach, we must select the model order *n*, then use some technique to estimate the model parameters, and finally validate the model. Usually, the model order is chosen using engineering judgment and some experimentation. One of main applications for AR modeling is in spectral estimation where it is necessary to estimate the shape of the spectrum of the signal and, in particular, to estimate the frequencies present in the signal. Then validation would mean that the frequencies have been estimated correctly within some tolerance. For the present application validation of a model meant correct classification in the parameter feature space.

Modeling of the time-series is like compression since usually the number of signal samples is much larger than the number of parameters. Therefore, for automatic fault classification it is preferable to use the AR parameters. However, typically to get a good

match between the signal and the model or to obtain a good spectral estimate a large model order is necessary. For example, the number of parameters is typically more than twice the number of frequencies present in the signal. For classification, it is not necessary to represent all the features of the signal; rather, it is necessary to correctly estimate those features that help in distinguishing among the different signal types.

A variety of techniques can be applied for classification, such as nearest neighbor techniques, or Fisher's linear discriminant method, etc. [2]. The representation of a signal or its model fixes the space in which the signal is represented as a point. The goal of pattern classification is to distinguish among the different points arising from the different types of signals. The points in feature space can be separated with boundaries often called discriminants. If the signals can be represented in a lower dimensional space without compromising classification, then the complexity of those boundaries is usually simpler. Dimension reduction usually identifies the principal components of the feature vectors. This also provides further compression without loss of classification performance but leads to simpler classifiers.

Fault Classification with Reduced Dimension AR Parameters:

We have developed a new approach for machinery fault classification through AR modeling followed by dimension reduction, utilizing techniques prevalent in the statistics and pattern recognition areas. The goal is to enhance those features of the AR-coefficient vector that contribute the most to the classification of faults. We describe the details of our hybrid approach and illustrate it by analyzing the Westland helicopter transmission vibration signal data set. The data consist of 8 accelerometer signals for each of 68 no-fault and seeded-fault runs on the aft transmission of a CH-46E helicopter. The data are highly oversampled (Figure 1 & Figure 2) and exhibit stationarity (Figure 3), which permits us to split each data file into multiple 1-second segments and decimate the data. This reduces processing time and provides for additional training data, allowing us to obtain statistically significant results. Here, the fault numbers refer to the following fault conditions: input pinion bearing corrosion (3), spiral bevel input pinion spalling (4), helical input pinion chipping (5), helical idler gear crack (6), collector gear crack (7), quill shaft crack (8), no defect (9). Accelerometer number 3 (aft side of mix box) was used for most of these results.

Processing Flow of the New Hybrid Approach:

Before presenting the results of the new approach we will describe the steps that are performed. The processing flow is shown in Figure 4. Notch filtering removes the dominant mesh frequencies to allow for better modeling of the other features present in the time series. We used the System Identification toolbox of MATLAB to perform the AR modeling. There are many dimension reduction techniques available and we selected the Fisher's feature selection method and a non-parametric method that were available in the SEPARAT toolbox, a MATLAB-based collection of mostly statistical pattern

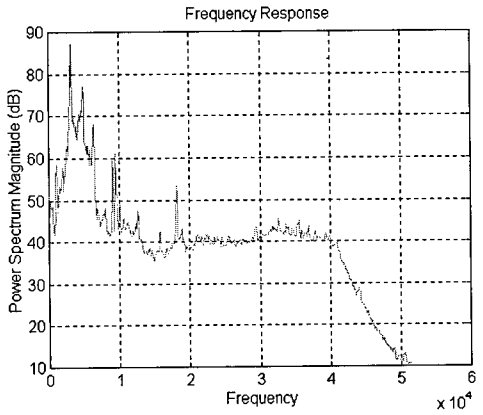


Figure 2: Power Spectrum of an accelerometer time series in an input bearing corrosion run.

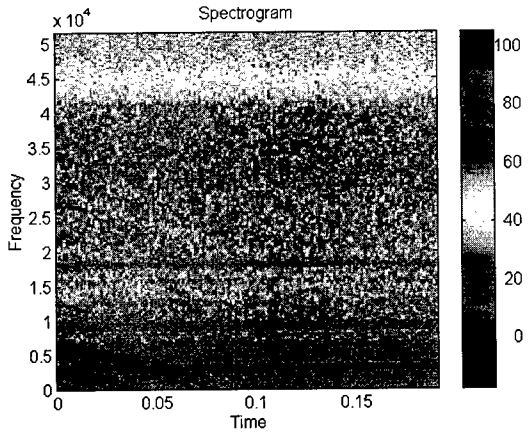


Figure 3: Spectrogram during the same run as in Figure 2.

classification techniques that was developed at the Penn State Applied Research Laboratory. Also used from SEPARAT were the pattern classification techniques, viz., the Fisher's Linear Discriminant and the k Nearest Neighbor (k-NN) methods. Typically, we developed an AR model of order 50 or 20 or 30 and then reduced the dimension down to 2. This dramatic reduction in features is possible since our only goal is to be able to distinguish among the different types of time series.

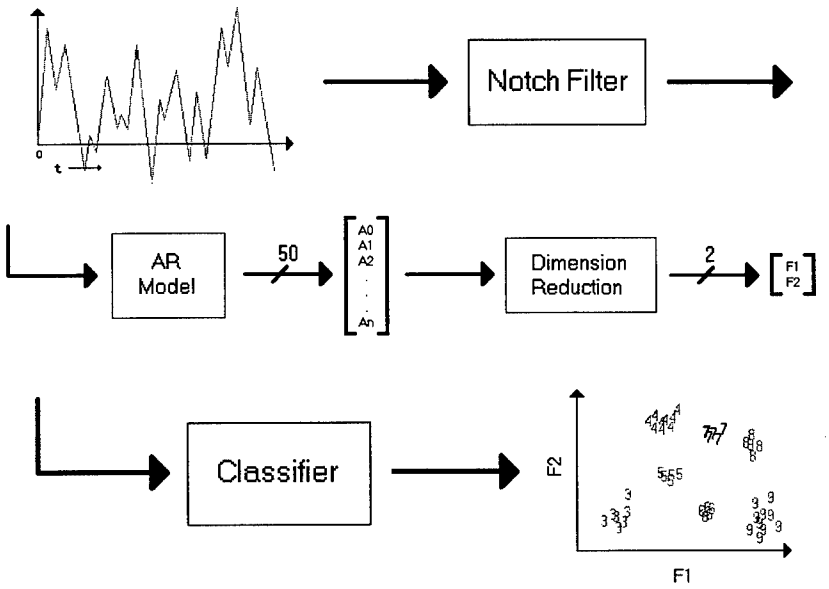
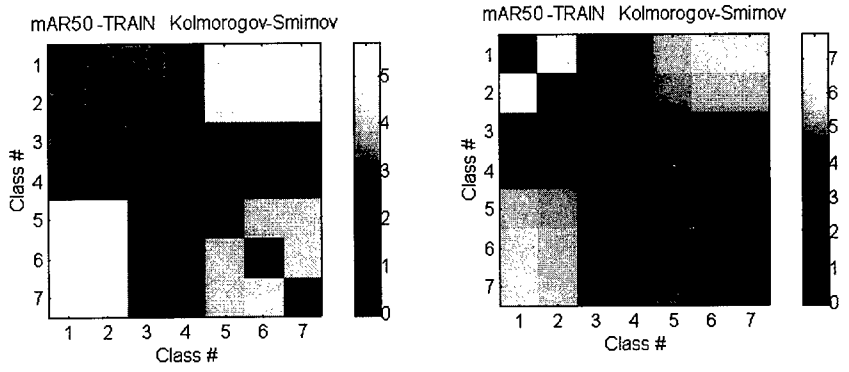


Figure 4. Processing flow for the novel fault classification approach.

Results of Data Analysis:

We will demonstrate our findings with the use of a separability measure. Briefly, a separability measure is formed to quantify the ease with which the patterns or points from different classes (or categories) can be distinguished. The Kolmogorov-Smirnoff (K-S) and Fisher's separability measures are often used for this purpose [2] and are available in SEPARAT toolbox. We chose to use the K-S measure since it allowed us to better characterize the different approaches and their relative advantages. The results are displayed as m x m matrices (where m is the number of classes or faults) in which the intensity of the element $z(I,J)$ shows the separability measure between classes I and J. The lighter the color the higher the separability. It is important to note also the scale on these matrices which varies very greatly from technique to technique.

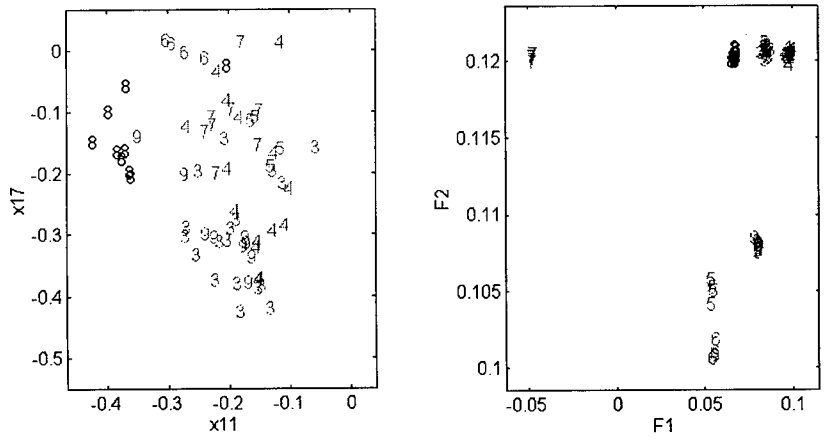
Classification performance is very dependent on the accelerometer used for generating features; as expected, the accelerometer closest to the faulty component provides the most information about the fault. This result is shown in Figure 5.



(a) (b)

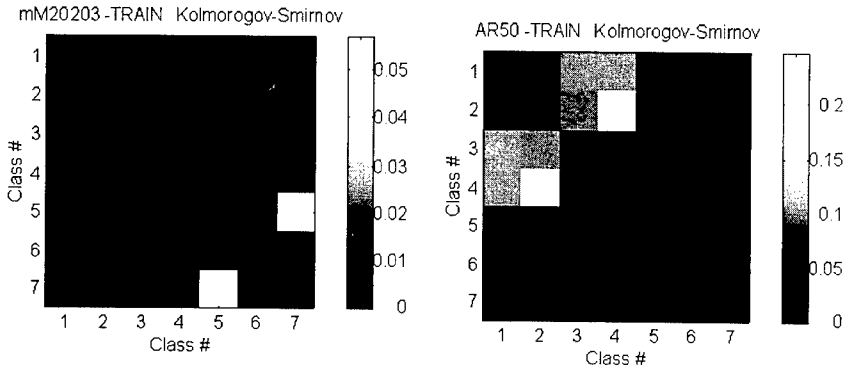
Figure 5. Separability results for 50-AR model reduced to 2-feautre space via Fisher's method: (a) Accelerometer 1 and (b) Accelerometer 3.

Based on the test rig schematic and on initial exploration we selected Accelerometer 3 (aft side of mix box) for most of our remaining analysis. Direct AR modeling of the signals requires high-order models due to the presence of many mesh frequencies and corruptive noise. However, such high-order models contain redundant information which does not help in fault classification and, in fact, leads to unreasonable training data size requirements, that are both cumbersome and expensive. Consequently, the AR coefficient feature vectors without post-processing result in poor separability in feature space. This is demonstrated by Figure 6 where Fisher's reduction is used for dimension in Figure 6 (b). In Figure 7, we show that it is important to model with high enough model order.



(a) (b)

Figure 6. Clusters for (a) 50-coefficient and (b) 2-feature vectors.



(a) (b)

Figure 7. Separability with (a) 20-AR and 50-AR feature vectors (no dimension reduction performed).

We compared the effect of different dimension-reduction techniques on fault classification. We show that for higher-order AR vectors all the dimension reduction techniques give comparable results. But, for lower-order AR vectors non-parametric methods result in greater fault separability at the cost of greater computation. It must be pointed out that the reduction is not simply sorting coefficients by their significance to classification and then retaining the top two or three. Rather, the reduction, in the case of the Fisher method, performs a rotation of the model space in order to find the principal components of the model space. The other reduction technique employs a non-linear algorithm to perform the same operation. An added benefit being accrued from this research is a better understanding of the manifestation of fault modes and the way they exhibit themselves in observable signals. The results of the comparison using various combinations of reduction and classification techniques are shown in Table 1 where the classification performance is provided.

Finally, In Figure ? we have shown the classification boundaries that result from using the Fisher's feature reduction and Fisher's discriminant mehhod on 50-AR coefficient vectors derived from the original time series. It is quite clear that there is excellent separation in the clusters, which is the reason behind the perfect classification performance. For testing purposes we separated the data files into two roughly equal sets (about 300 data files each) and used one set for training and the other for testing.

Table 1. Classification Results with Various Reduction & Classification Techniques.

Technique	*% Correct*
50-AR / FR / FD	100 %
50-AR / FR / 2-NN	100 %
50-AR / NLR / 2-NN	100 %
50-AR / X / FD	94 %
50-AR / X / 2-NN	92 %
20-AR / X / FD	40 %
20-AR / FR /FD	58 %
20-AR / X / 2-NN	26 %
20-AR / FR / 2-NN	38 %

Notes: AR = Autoregressive; FR = Fisher Reduction; NLR = Non Linear Reduction; X = No Reduction; 2-NN = Two nearest neighbor classification; FD = Fisher Discriminant Classification.

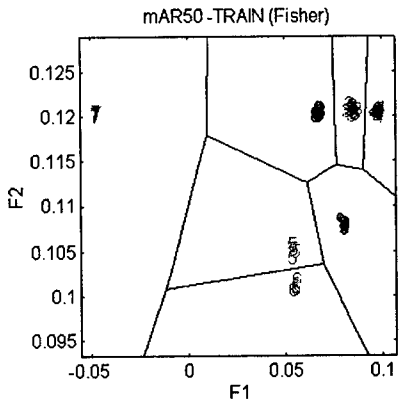


Figure 8. Classification Boundaries for 50-AR features reduced to 2-feature vectors with Fisher reduction and using Fisher Discriminants.

Conclusions and Future Research:

We have described a new approach for efficient and high performance fault classification using standard AR modeling followed by dimension reduction. The signal is first modeled with a 50-coefficient AR model, then the dimension is reduced to 2, and the feature vectors are classified in 2-feature space. This approach results in reduced complexity classifiers, reduced training data requirements, and greater understanding of the operation of the fault classification process. This observation and the fact that the resulting classifiers all involve linear operations should make this technique very suitable

for real-time applications. We utilize various dimension reduction and linear and nonlinear classifier methods. The approach achieves near perfect classification performance in blind testing. The data used were collected by Westland Helicopters on the aft transmission of the CH-46E helicopter. Most pattern recognition algorithms (even the linear ones) could perform the classification with an accuracy of greater than 90%.

In closing, several areas of future research are identified that are relevant to the areas of CBM and fault monitoring/tracking: neural network based classification, nonlinear regression modeling techniques, the implication of the transformations on the utility of the sources of information, and various other extension and outgrowth areas.

References:

[1] Baillie, D. C. and J. Mathew, "A Comparision of Autoregressive Modeling Techniques for Fault Diagnosis of Rolling Element Bearing," *Mechanical Systems and Signal Processing*, Vol. 10(1), 1-17, 1996.

[2] Duda, R. O. and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, 1973.

[3] Garga, A. K., "Hybrid Reasoning Techniques for Automated Fault Classification," Proceedings of the 51st Meeting of the Mechanical Failures Prevention Group, Va. Beach, VA, April 15-17, 1997.

[4] Garga, A. K., "A Hybrid Implicit/Explicit Automated Reasoning Approach for Condition-Based Maintenance," *ASNE Intelligent Ships Symposium II*, Philadelphia, PA, November 25-26, 1996.

[5] Hall D. L., "Concepts of condition-based maintenance for intelligent ships," *ASNE Intelligent Ships Symposium II*, Philadelphia, PA, November 25-26, 1996.

[6] Hall, D. L., *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Norwood, MA, 1992.

[7] Hansen, R. J., D. L. Hall, and S. K. Kurtz, "A new approach to the challenge of machinery prognostics," *International Gas Turbine and Aeroengine Congress and Exposition*, Netherlands, June 13-16, 1994. (Also in the *Transactions of the ASME*, **117**:320-325, April 1995.)

[8] Lim, J.S. and A. V. Oppenheim (Eds.), *Advanced Topics in Signal Processing*, Pretice Hall, 1988,

[9] Stadter, P. A. and A. K. Garga, "A Neural Architecture for Fuzzy Classification with Application to Complex System Tracking," in review for the *International Conference on Neural Networks*, 1997.

Acknowledgments:

This work was supported by a grant from the Office of Naval Technology (the Multi-Disciplinary University Research Initiative for Integrated Predictive Diagnostics Grant no. N00014-95-1-10461). We would like to thank Dr. Greg Babich for providing the beta version of his SEPARAT software.

MODEL-BASED TRACKING OF HIDDEN PARAMETER DRIFT IN DYNAMICAL SYSTEMS

J. P. Cusumano, D. Chelidze, and N. Hecht
Department of Engineering Science & Mechanics
Penn State University
University Park, PA 16802

Abstract: A model-based experimental method for tracking parameter drift in nonlinear dynamical systems is described. Local linear tracking models are constructed using data sampled over a fast time scale. These models are used to analyze data from systems with parameters which evolve over a slow time scale according to a "hidden" rate law. The method is applied to a numerical study of a nonlinear electrical circuit with a variable resistance as the drifting parameter. The mean-square tracking model prediction error is shown to follow successfully both ramped and sinusoidal parameter variations, suggesting that, at least in the cases studied, the method provides an invertible mapping between the parameter space and the observable space. Thus it should be possible to extract *rate information* about hidden drift, a requirement for true prediction.

Key Words: Condition monitoring; dynamical system; model-based; parameter drift; prediction.

INTRODUCTION: In this paper, a model-based method for tracking parameters which are drifting due to some "hidden" rate law is described. The aim is to develop a method which is capable of monitoring hidden processes such as those which are typical in many dynamical systems with evolving damage. In general, damage processes are hidden from the observer. Worse, an appropriate damage state variable is usually not known. However, one can observe the performance of the machinery or electronic equipment, since one can measure physical quantities such as temperature, displacement, velocity, force, voltage, etc. This paper demonstrates that it is possible to track the state of a hidden process by analyzing the available observable quantities, even when direct access to the drifting parameters is not available.

The method discussed here is suitable for monitoring systems for which the hidden processes evolve on a much slower time scale than the observable dynamics. This assumption of time-scale separation may not always be appropriate, but it can be expected to hold in a wide variety of applications. For example, fatigue damage in a structure might accumulate over thousands of hours of operation, but the vibrational response of the structure can be characterized and considered to be approximately stationary over a time scale of seconds or minutes. If "damage" is interpreted broadly to include *any* unwanted change in system properties, the same can be said of the

gradual drifting out of alignment of components in a mechanical or electromechanical system over time as the result of ambient vibrations. Finally, electronic systems which operate in the kilohertz or higher range but in which, e.g., contact resistances or supply voltages gradually change over the course of months or years, fall into the same category.

The idea of using model-based techniques to identify certain irregularities in dynamical systems is certainly not new. Fault diagnosis in dynamical systems has been studied by many researchers (see, for example, [1], and the references therein). Failure detection and isolation (FDI) theory is one approach which has received considerable attention, with application, for example, to the detection of sensor failure [2]. The FDI method uses a known relationship between the inputs and the outputs of a system to anticipate the system response for any given input. Any change in the system will produce a discrepancy between actual and FDI predicted output values. The main disadvantage of this theory is that it requires an analytical system model (i.e. known relationship between input and output). Unfortunately, in many cases the analytical system model is unknown, or is hard to determine, especially for nonlinear systems [3].

Given that most damage or wear processes are fundamentally nonlinear, an approach that can accomodate nonlinearity in the underlying system is desirable. It is well known that many essential characteristics of nonlinear dynamical systems can be extracted from even a single scalar time series using delay coordinate embedding [4, 5, 6, 7, 8]. State-space modeling techniques based on delay coordinate embedding do not require knowledge of governing equations, and are solely based on topologically invariant features of the underlying systems. While a great amount of work has been done in the area of nonlinear predictive models constructed from data (as summarized, e.g. in [9, 10]), there has been a heavy emphasis on studying the fundamental features of chaotic systems, which is not our concern here. In this paper it is shown that even the simplest class of such models (those which are *locally linear*) capture enough information from underlying dynamics to make it possible to track changes in hidden parameters.

DESCRIPTION OF METHOD: We consider hierarchical systems of the form:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \beta, t), \tag{1a}$$

$$\dot{\beta} = \epsilon\, g(\mathbf{x}, \beta, t) \tag{1b}$$

where $\mathbf{x}$ is the observable state vector, and β is the drifting parameter, here assumed to be a scalar and "hidden" (i.e. not directly observable). The rate constant ϵ is assumed small (i.e. $\epsilon \ll 1$). The vector field $\mathbf{f}$ can be thought of as approximately governing the dynamics of the observable state $\mathbf{x}$, however the subsystem (1a) is nonstationary on the slow time scale $t = \mathrm{O}(1/\epsilon)$ due to the hidden rate law (1b). Note that when $\epsilon = 0$, the parameter β is fixed, and thus acts as a true parameter in (1a). The system thus possesses two distinct time scales. Therefore, during data acquisition at the fast time scale $t = \mathrm{O}(1)$ the parameter β may be treated as approximately con-

stant, and tracking models may be constructed that describe the subsystem (1a) at an fixed value of β. In this way, a sequence of tracking models can be obtained that, as is demonstrated below, can be used to monitor variations in β over slow time scales.

Tracking is accomplished using local linear models computed from short time duration samples of fast time data (that is, data sampled at the time scale of O(1)). The first step in the process is to perform a delay coordinate embedding (or "time delay reconstruction") of the original time series data. Given an original scalar time series $\{x(n)\}_{n=1}^{M}$, where M is the total number of data points in the sample, one creates d-dimensional column vectors $\mathbf{y}(n)$ of the form:

$$\mathbf{y}^T(n) = (x_1(n), x_2(n), \ldots, x_d(n)) \tag{2}$$

where $x_k(n) = x(n+(k-1)\tau)$, d is the embedding dimension, and τ is the time delay. A suitable time delay can be obtained by finding the first minimum of the average mutual information [11], and the method of false nearest neighbors can be used to determine the minimum required embedding dimension [12].

Working in the global reconstructed phase space of the dynamical system, local models which show how neighborhoods about each data point are mapped forward in time can be estimated. The simplest such model is the local linear model relating the state at time n, $\mathbf{y}(n)$ to the state $\mathbf{y}(n+1)$ at time $n+1$:

$$\mathbf{y}(n+1) = \mathbf{A}_n\mathbf{y}(n) + \mathbf{a}_n = \mathbf{B}_n\hat{\mathbf{y}}(n), \tag{3}$$

where the model parameter matrix $\mathbf{A_n}$ and parameter vector $\mathbf{a_n}$ are determined by regression at each point in the data set, and where

$$\hat{\mathbf{y}}(n) = \left[\mathbf{y}^T(n), 1\right]^T, \quad \text{and} \quad \mathbf{B}_n = [\mathbf{A}_n\ \mathbf{a}_n] = \left[b_{pq}^n\right]. \tag{4}$$

Basing our model on N nearest neighbors ($\mathbf{y}^r(k)$, $r = 1, \ldots, N$) of the point of interest ($\mathbf{y}(k)$) we can determine the optimum model parameters by minimizing

$$\mu = \sum_{r=1}^{N} |\mathbf{y}^r(k+1) - \mathbf{B}_k\hat{\mathbf{y}}^r(k)|^2 = \sum_{r=1}^{N}\sum_{n=1}^{d}\left(y_n^r(k+1) - \sum_{m=1}^{d+1} b_{nm}^k\hat{y}_m^r(k)\right)^2. \tag{5}$$

Varying μ with respect to each b_{pq}^k (with k fixed) we obtain

$$\sum_{r=1}^{N} y_p^r(k+1)\,\hat{y}_q^r(k) = \sum_{r=1}^{N}\sum_{m=1}^{d+1} b_{pm}^k\hat{y}_m^r(k)\,\hat{y}_q^r(k). \tag{6}$$

It is then easy to show that,

$$\mathbf{B}_k = \mathbf{Y}_{k+1}\hat{\mathbf{Y}}_k^T\left(\hat{\mathbf{Y}}_k\hat{\mathbf{Y}}_k^T\right)^{-1}, \tag{7}$$

where

$$\mathbf{Y}_k = \left[\mathbf{y}^1(k)\,\mathbf{y}^2(k)\,\ldots\,\mathbf{y}^N(k)\right]\ , \quad \hat{\mathbf{Y}}_k = \left[\hat{\mathbf{y}}^1(k)\,\hat{\mathbf{y}}^2(k)\,\ldots\,\hat{\mathbf{y}}^N(k)\right]. \tag{8}$$

In utilizing the above formulas to construct the tracking models, the *k-d* tree algorithm [13, 14] is used to reduce the calculation time required for finding the nearest neighbors about each point. Although it is outside of the scope of this paper, we remark that state-space models similar to those described above can be enhanced perform various filtering operations, such as cleaning noise-corrupted signals [7, 8].

The main assumption in using models such as that defined by equation (3) to track parameter drift is that models based on reconstructed phase space data capture enough information about the underlying system at the moment of data acquisition to provide a sort of "snap shot" of the system. By comparing the model performance, as measured with some suitable metric, for models estimated from data sets collected at different times, one can attempt to observe changes in the system over the slow time scale $O(1/\epsilon)$. For this paper, the root mean square (RMS) single step prediction error e_1 was used as the tracking model performance metric. In other words, e_1 serves as a *drift observable.*

APPLICATION TO A NONLINEAR CIRCUIT: The method described in the previous section was applied in numerical experiments conducted with a model of a nonlinear circuit due to Matsumoto and co-workers [15, 16]. The circuit is easily implemented in hardware, exhibits a wide range of periodic and chaotic behaviors, and is modeled with reasonable accuracy by a well-understood system of differential equations. It is thus well-suited to combined experimental, analytical, and numerical studies.

As shown in [15, 16], the circuit can be described by the following dimensionless system of equations:

$$\begin{aligned} \dot{x} &= \alpha\,(\,y - h\,(x)\,)\,, \\ \dot{y} &= x - y + z, \\ \dot{z} &= -\,\beta\,y, \end{aligned} \tag{9}$$

where the state variables $(\,x, y, z\,)$ represent voltages and the parameter β represents a resistor that controls the behavior of the system. The piecewise linear function $h(x)$ is a nonlinear resistance that has the form

$$h(x) = m_1\,x + (\,m_0 - m_1)\,(\,|x+1| - |x-1|\,)\,/2 \tag{10}$$

for constants m_0 and m_1. To the original system (9), we add a rate law that causes a drift in the control resistor β, with the general form:

$$\dot{\beta} = \epsilon\, f(\,x, y, z, \beta, t\,). \tag{11}$$

Following Matsumoto [15], we set $\alpha = 7$, $m_0 = -1/7$, and $m_1 = 2/7$. The system is known to be chaotic for $\beta \in [6.5, 10.5]$: after obtaining a bifurcation diagram numerically, the range $\beta \in [9.2, 9.3]$ was chosen for further numerical study. A small range was selected deliberately so that the power spectra and probability density functions for the steady state response of the system *did not* display substantial change while β was varied from 9.2 to 9.3.

Three separate simulations were carried out corresponding to three different choices for the rate law (11). In all three cases, the local linear model was constructed for $\beta = 9.2$, and then the RMS single step prediction error e_1 for the model was computed using data obtained for values of β throughout the interval. In the first case, equation (11) is just

$$\dot{\beta} = 0 \tag{12}$$

and thus β is constant. This case serves as a control since β is strictly constant during the data acquisition phase of the simulations. The second case used equation (11) in the form

$$\dot{\beta} = \epsilon. \tag{13}$$

The form of rate law in equation (13) corresponds to a ramped parameter drift $\beta = \epsilon t + \beta_0$ over the slow time scale. The third and final case studied took the rate (11) with the form

$$\dot{\beta} = \epsilon B \sin(\epsilon t), \tag{14}$$

which, of course, corresponds to a cosinusoidal variation in β.

For the fixed β simulation corresponding to the rate law (12), 21 equally-spaced values of β in the interval $[9.2, 9.3]$ were selected as $\beta_m = 9.2 + 0.005\,m$ for $(m = 0, \ldots, 20)$. For each β_m the system (9) was run to steady state using a variable step size Runge-Kutta algorithm starting from the initial conditions $[0.01, 0.01, 0.01]$. Then data sets of 80,000 points were collected at a sample rate of 10 (samples)/(time unit). For reference, the power spectral density was approximately 90dB down from its peak value at a frequency on the order of 1 (in dimensionless frequency units). Thus, our data was sampled in excess of 5 times that required by the Nyquist criterion.

For the ramped case of equation (13) the simulation started by running to steady state with the initial value of $\beta = 9.2$. Then the simulation continued using equations (9) and (13). Again 80,000 data points were collected at a sample rate of 10 starting at the same 21 β_m used for the fixed β case; however, in this case β continued to drift throughout the data acquisition phase. The parameter $\epsilon = 6.1881 \times 10^{-9}$ was chosen so that the total data collection time (8,000 time units) was $1/100^{\text{th}}$ of the time between β_m's.

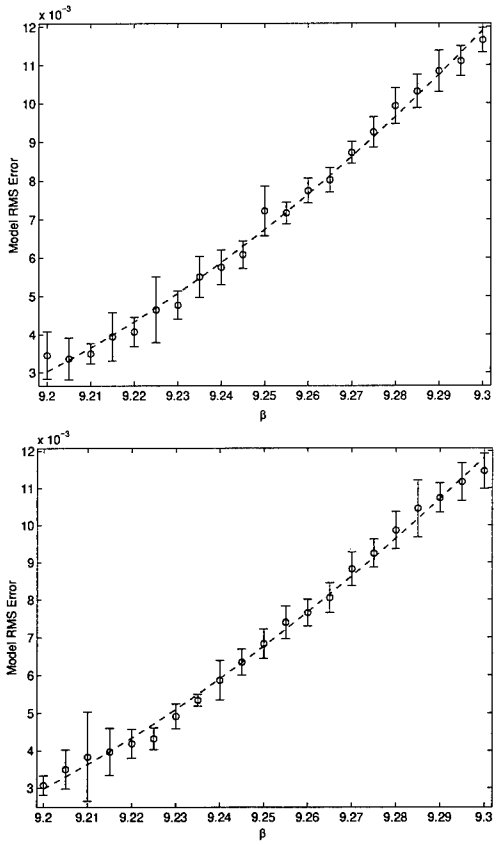


Fig. 1: Raw plots of RMS single step error e_1 vs. β for (top) the constant β case (equation 12); and (bottom) the ramped case (equation 13). Error bars were obtained by computing e_1 for 10 different input data sets. The dashed line represents a quadratic fit.

For the cycled parameter "drift" simulation corresponding to (14), the procedure followed closely that used for the ramped case. The amplitude was set to $B = 0.05$ to cover the same range of β used for the other cases. Hence, data was collected starting at each $\beta_m = 9.25 - 0.05\cos(2\pi m/20)$ for $(m = 0, \ldots, 20)$. The rate constant was set to $\epsilon = 3.8881 \times 10^{-7}$ so that, as in the ramped case, the ratio between the data acquisition time starting at a given β_m and the waiting time between successive β_m's was 1/100.

Average mutual information and false nearest neighbors algorithms showed that a delay parameter τ of 7 or 8 with an embedding dimension d of 4 would give sat-

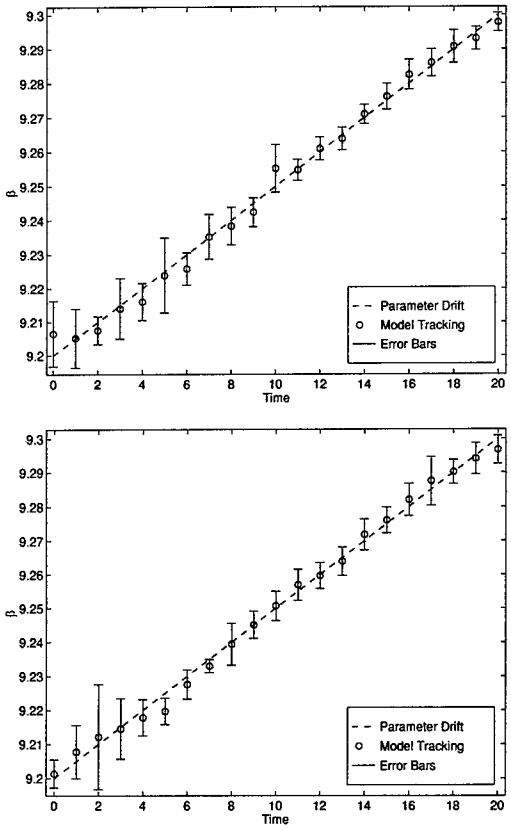


Fig. 2: Tracking results using the calibrated e_1 vs. β curves from : (top) static case; and (bottom) ramped case. Note that the horizontal "time" axis corresponds to the data set number for the fixed β case, whereas it represents a true continuous time in the ramped case.

isfactory results for phase space reconstruction based on one of the state variables. In particular, only the state variable x was used for state space reconstruction and modeling using equation (3). The local linear model was constructed based on the first 2^{14} data points of the first scalar data set (i.e. starting with $\beta = 9.2$). The model was then applied as a predictor to all 21 data sets in each simulation. Each data set was divided into 10 different 2^{13} point data subsets, and simulations were performed on each subset. The mean value of the RMS predictor error e_1 was computed over the 10 subsets, along with standard deviations of the mean to be used as an estimate of the expected fluctuation in e_1.

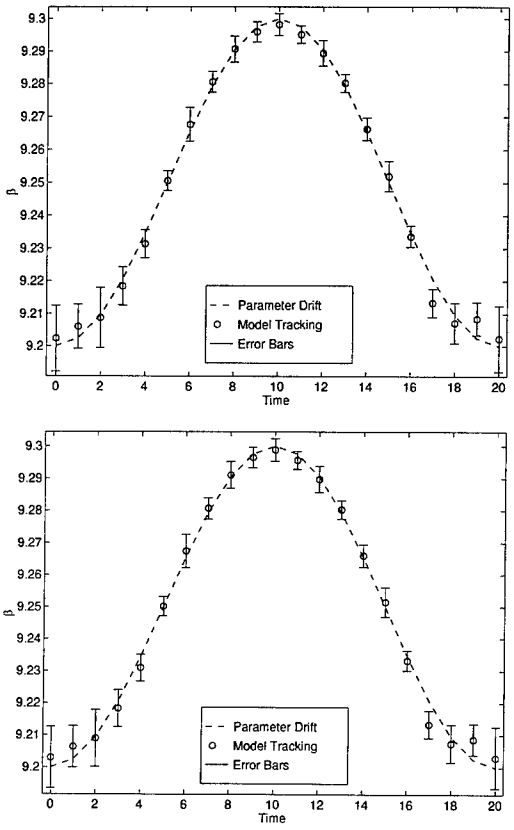


Fig. 3: Tracking results using the calibrated e_1 vs. β curves from Fig. 1 applied to the cosinusoidal β variation (equation 14): (top) calibration with static result; and (bottom) calibration with ramped result. For both calibrations, the drift observable e_1 tracks β to within experimental fluctuations.

RESULTS: Plots of raw (unscaled) results showing e_1 vs. β are presented in Fig. 1 for the fixed and ramped cases. Examination of the figures indicates that there is a nonlinear characteristic relating the drift observable e_1 and β. Note that, in both cases, as $\beta \rightarrow 9.2$, e_1 does not go to zero but rather goes to the residual predictor error of the model generated with the original $\beta = 9.2$ data set.

Using the raw data, a polynomial fit was used calibrate the drift observable e_1 to the known range of β: a quadratic polynomial was found to fit the results to within experimental error. The results of transforming the raw data in Fig. 1 using the resulting fits are shown in Fig. 2. The observable e_1 tracks β to within experimental

fluctuations for both the static control case and with the ramp drift.

A similar calibration to that used for Fig. 2 could be carried out with the cosinusoidal drift (equation (14)), however a more interesting idea is to use the calibration from one simulation to attempt tracking with the data from a completely different simulation. In Fig. 3 , the results of using calibrations obtained with the fixed and ramped simulations to track the cosinusoidal drift case are shown. This provides a check of the internal consistency of the results, and suggests that the mapping $e_1(\beta)$ from the parameter space to the observable space does not depend strongly on the specific form of the drift, a result which would be important in applications.

DISCUSSION AND CONCLUSIONS: Using a nonlinear model constructed from data for a system at some initial time, we have shown that the RMS single step predictor error e_1 serves as a drift observable: that is, it can be used to track slow changes in a hidden parameter. For sufficiently slow drift, it was demonstrated that tracking models constructed over sufficiently short times are not adversely affected by the nonstationarity present during data acquisition. By successfully tracking both ramped and cosinusoidal parameter variations, it was shown that, at least for the conditions of the numerical experiments, the mapping between the drift observable and the hidden parameter is 1-1 (invertible). Thus, using this technique it should be possible, at least in principle, to extract *rate information* about the drift, a necessary requirement for true prediction of system parameter changes.

Several different experimental applications of the method are currently underway. In particular, a hardware implementation of the system studied here has been built with a digitally-controlled resistor. Issues which will be explored in future work include: the effect of time scale separation on performance (in the case studied here, a 1:100 separation was imposed); the rate at which the tracking model needs to be updated; and noise tolerance.

ACKNOWLEDGEMENTS: This work is supported by the Office of Naval Research MURI on Integrated Predictive Diagnostics, Grant #N0014-95-0461.

REFERENCES

[1] Patton, R., Frank, P., and Clark, P. (1989) *Fault Diagnosis in Dynamical Systems: Theory and Application*, Prinston Hall International (UK) Ltd.

[2] Rizzoni, G., and Min, P. S. (1991) "Detection of Sensor Failure in Automotive Engines" *IEEE Transaction on Vehicular Technology*, **40**(2), pp. 487–500.

[3] Ribbens, W. B. (1992) "Distinguishing Changes in Plant Dynamics from Actuator or Sensor Failures for Failure Detection in Nonlinear Plants" *ASME Transactions on Transportation Systems*, DSC- **44**, pp. 39–52.

[4] Packard, N. H., Crutchfield, J. P., Farmer, J. D., and Shaw, R. S. (1980) "Geometry from a Time Series" *Physical Review Letters*, **45**, pp. 712–716.

[5] Takens, F. (1981) "Detecting Strange Attractor in Turbulence" in *Dynamical Systems and Turbulence*, D. A. Rand and L. S. Young, eds., Springer Lecture Notes in Mathematics, **898**, pp. 266–281, Springer-Verlag: New York.

[6] Sauer, T., Yorke, J. A., and Casdagli, M. (1991) "Embedology" *Journal of Statistical Physics*, **65**(3,4), pp. 579–616.

[7] Casdagli, M., Eubank, S., and Farmer, J. D. (1991) "State Space Reconstraction in the Presence of Noise" *Physica D*, **51**, pp. 52–98.

[8] Sauer, T. (1993) "Time Series Prediction Using Delay Coordinate Embedding" in *Time Series Prediction: Forcasting the Future and Understanding the Past*, A. S. Weigand and N. A. Gershenfeld, eds., Santa Fe Institute Studies in the Sciences of Complexity, Proceedings Vol. XV, Addison-Wesley: Reading, MA, pp. 175–193.

[9] *Time Series Prediction: Forcasting the Future and Understanding the Past*, A. S. Weigand and N. A. Gershenfeld, eds., Santa Fe Institute Studies in the Sciences of Complexity, Proceedings Vol. XV, Addison-Wesley: Reading, MA.

[10] Abarbanel, H. D. I. (1995) *Analysis of Observed Chaotic Data*, Spring Verlag.

[11] Fraser, A. M., and Swinney, H. L. (1986) "Independent Coordinates for Strange Attractors from Mutual Information" *Physical Review A*, **33**(2), pp. 1134–1140.

[12] Kennel, M. B., Brown, R., and Abarbanel H. D. I. (1992) "Determining Embedding Dimmension for Phase-Space Reconstruction Using a Geometric Construction" *Physical Review A*, **45**(6), pp. 3403–3411.

[13] Friedman, J. H., Bentley, J. L., and Finkel R. A. (1977) " An Algorithm for finding Best Matches in Logarithmic Expected Time" *ACM Transactions on Mathematical Software*,**3**(3), pp. 209–226.

[14] Bentley, J. L., and Friedman, J. H. (1979) "Data Structures for Range Searching" *Computing Surveys*, **11**(4), pp. 397–409.

[15] Matsumoto, T. (1987) "Chaos in an Electrical Cicuit" *Proceedings of the IEEE*, **75**(8), pp. 1033–1057.

[16] Matsumoto, T., Chua L. O., and Ayaki, K. (1988) "Reality of Chaos in in the Double Scroll Cicuit: A Computer-Assisted Proof" *IEEE Transaction on Circuits and Systems*, **35**(7), pp. 909–925.

The Algebraic Determination of Current Components in Induction Motor Driven Machinery

Reginald G. Mitchiner and Michael Kaiser

Department of Mechanical Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, 24061-0238 USA

Abstract: This paper derives expressions for and discusses the three phase current components, in the frequency domain, in an induction motor which drives a mechanical load. The expressions presented illustrate the dependence of the various current components on the dynamics of the rotor (angular acceleration, velocity, and position), on the motor electrical parameters, and on the mechanical load parameters.

The mechanical loads imposed on the motor include harmonics of the mechanical load torque. Thus the perturbations of the currents, in the frequency domain, are linked to perturbations of the harmonic components in the imposed load function.

The basis for the derivations are the motor equations of Stanley and Kliman. It is assumed that the dynamics of the motor rotor are available through direct measurement. Using experimental measurements and simulations, the algebraic relationships are confirmed.

The utility of the relationships derived herein is that one may appropriately choose a set of motor parameters responsive to specific variations of mechanical parameters in the detected current components. Further, one may easily identify both those current components which are responsive to changes in mechanical parameters and those which are not responsive. Thus in the area of mechanical diagnostics using motor currents and parameters, one may design and utilize a system with quantitative expectations.

Key Words: AC motor; current perturbations; induction motor; phase currents.

Introduction: Many of the failures of electric induction motor driven rotating machinery are due to the applied mechanical loads. These failures impose upon the rotor a torque signature in the time and frequency domains that is different from that observed in normal (failure-free) operation.

Common ways to measure these torques are by attaching strain gages to the shaft or by inserting a special section of shafting which acts as a transducer. This type of shafting is usually expensive and more often not possible to use. If an easily implemented method were available to monitor the mechanical load (or torque) over a period of time, such things as piston, bearing, and pump impeller wear (among others) could be

detected before a total failure occurred.

In this paper, the mechanical loads imposed are pure torque harmonics which simulate loads imposed on the driving electric motor with a direct dependence on the position, velocity, and acceleration of the reciprocating element. Thus the perturbations of the phase currents are dependent on the perturbations of the load parameters in the machine.

It is the purpose of this paper to identify these perturbations and to quantify the load parameters through identification of explicit algebraic expressions for the phase currents and the load torque. Accomplishment of this requires sensing the phase currents and voltages of the motor, as well as the position, velocity, and acceleration of the rotor. The rotor dynamics are assumed to be available through direct measurement. This paper presents expressions for the three phase current components, in the frequency domain, of an induction motor which drives a time varying mechanical load.

Induction Machine Equations: The following section is presented as a review of the three phase machine equations, based on the work of H. C. Stanley[1], P.C. Krause and C.H. Thomas[2], and the work of G. Kliman[3].

In the 1930's, Stanley demonstrated that all induction machines can be analyzed by an equivalent two phase machine model by transforming the rotor variables to a reference frame fixed to the stator. In 1965, however, P.C. Krause and C.H. Thomas noted that all real transformations used in the analysis of induction machines can be described by one general transformation of variables. The reference frame associated with this general transformation of variables is known as an arbitrary reference frame. It is this reference frame on which Kliman bases his observations.

Kliman has demonstrated that the steady state electromagnetic torque of an induction motor can be calculated from knowledge of the terminal voltages, currents, and stator resistance [1]. Kliman has shown that the electromagnetic torque produced by an induction machine, for constant loading, may be calculated from:

$$T_{em} = \frac{3}{2} p l_m \left[i_s \cdot \Psi_s \right] = \frac{3}{2} p \left(i_{sq} \Psi_{sd} - i_{sd} \Psi_{sq} \right) \qquad (1)$$

where, p = number of pole pairs,
i_s = stator current,
and Ψ_s = stator flux.

Note that Equation (1) depends only on stator quantities. The equivalent two phase current components i_{sq} and i_{sd} can be measured using non-contacting current sensors whereas the stator flux linkages, and are derived quantities. Kliman has shown that the stator flux linkages are derived from the stator voltages as follows:

$$\Psi_s=\Psi_{sd}+j\Psi_{sq}=\int_0^t \left(v_s-R_s i_s\right)\,dt \tag{2}$$

where, v_s = stator voltage,
R_s = stator resistance,
and i_s = stator current.

The equivalent two phase voltages and currents in terms of a three phase induction machine are as follows:

$$v_{sd}=v_a \tag{3}$$

$$v_{sq}=\frac{\sqrt{3}}{3}\left(v_b-v_c\right) \tag{4}$$

$$i_{sd}=i_a \tag{5}$$

$$i_{sq}=\frac{\sqrt{3}}{3}\left(i_b-i_c\right) \tag{6}$$

Note that the subscript (s) denotes stator, (a,b,c) denotes three phase quantities, and (d,q) denotes the equivalent two phase quantities.

Basis for Load Parameter Identification: By combining the electromagnetic torque demonstrated by Kliman with a rotary system's differential equation of motion, Equation (7), the load torque can be defined. More specifically, various load parameters can be identified.

$$J\ddot{\Theta}+C\dot{\Theta}+K\Theta=T_{em}-T_{Load} \tag{7}$$

where, J = total rotary inertia of system,
C = total rotary damping of system,
K = total rotary stiffness of system,
$\ddot{\Theta}$ = rotor angular acceleration,
$\dot{\Theta}$ = rotor angular velocity,
Θ = rotor angular position,
T_{em} = electromagnetic torque,
and T_{Load} = Load torque applied to rotor.

Assuming that damping and stiffness effects associated with the rotor are negligible, equation 7 reduces to equation 8.

$$J\ddot{\Theta}=T_{em}-T_{Load} \tag{8}$$

Combining Equation (1) with Equations (2) and (8) in the frequency domain, expressions for the stator currents in terms of motor quantities and load torque are formulated. This task was accomplished in the symbolic manipulation package DERIVE[4]. The results of the manipulations required to reduce the set of

simultaneous differential equations was protracted and Equation (9) shows a truncated sample of the results for one of the phase currents:

$$I_l = 12090*pi*(90420849552384000000000000*(I/S)**2*omega*(534251*S+34313448*pi)*(SQRT(3)*S+omega)**2+15475200000000*SQRT(3)*(1/S)*(7790602560000*SQRT(3)*((I5.8/(60*2*pi))**2*S+15.8/(60*2*pi)*(17.04)-I/(I/(15.8/(60*2*pi))+I/(17.57/(60*2*pi) \quad (9)$$

.....

$$29297442918400*pi**4*S**4+18267144384064000000000000000*S**2*(0.5*FLT/S)**2+2346495220153344000000000000000*pi*S*(0.5*FLT/S)**2+75354413673517056000000000000*pi**2*(0.5*FLT/S)**2)))$$

Note that Equation (9) contains motor constants which are defined below substituted in the symbolic expression. This was done to reduce the apparent size and complexity of the resulting expression.

Simulating a Load Torque: The expressions for the stator currents contain a simple variable which defines the load torque in the frequency domain. To illustrate the effects of mechanical load variation, the examples below show the results of Equation (9) with constant loads and harmonic loads comprised of various frequency components. These loads were imposed upon a commodity 1/2 horsepower Baldor 3 phase induction motor with the following characteristics:

Primary reactance	15.80	Ohms
Secondary reactance	17.57	Ohms
Primary resistance	17.04	Ohms
Secondary resistance	12.31	Ohms
Magnetizing reactance	183.80	Ohms
Rotor inertia	.90631E-2	Kg-m^2
Poles	8	
Phase voltage	120.9	Volts
Torque, full load	4.188	N-m

Effects of Load Parameters on Phase Currents: Figure 1 shows, in the frequency domain, the current spectrum derived from Equation (9) for the motor above when a constant 75% of the full-load torque is imposed upon the rotor. It should be noted that the magnitudes of the phase currents have been confirmed through time simulation of the Stanley equations and conversion of the time solutions into the frequency domain.

This spectrum, for the constant torque loading, show the expected broad-band characteristic with the magnitude of the current being dependent upon that of the imposed constant load.

Figure 2 illustrates the phase current spectrum when the imposed load torque is a

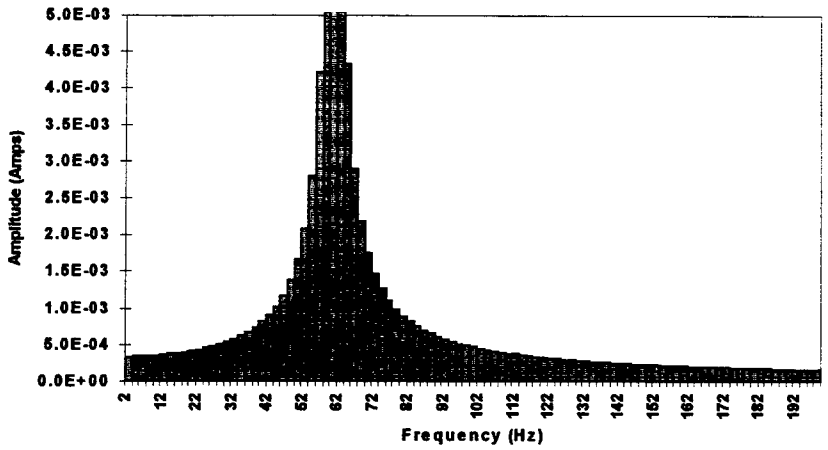


Figure 1: Fourier Transform of Phase Current with Constant 3/4 Full Load Torque Applied

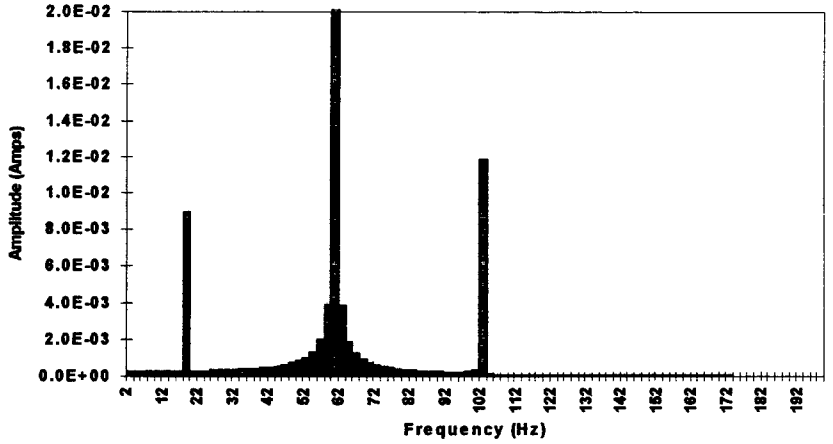


Figure 2: Fourier Transform of Phase Current with Constant 1/2 Full Load Torque Applied at 42 Hz Applied

single harmonic (of magnitude 1/2 FLT) at 42 Hz. The expected side band characteristic is displayed. The magnitude of the side bands is derivable from that magnitude of the imposed harmonic load.

When the imposed load includes a series of harmonics (in this case at 42 Hz, 84 Hz, and 126 Hz), additional side bands to the 60 Hz forcing frequency are created as shown in Figure 3. As before, the magnitude of the side band current components are derivable from those magnitudes of the original $T_{Load}(w)$ input into Equation (9).

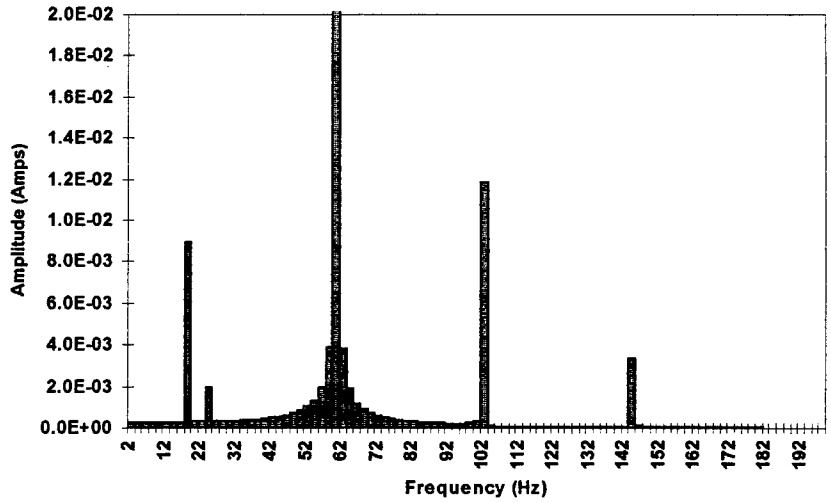


Figure 3: Fourier Transform of Phase Current with Constant 1/2 Full Load Torque Applied at 42, 84, and 126 Hz Applied

Closure: The authors have demonstrated herein that one, using a two axis model of the three phase induction motor, may derive an algebraic relationship, in the frequency domain, for the three phase stator currents. This relationship, while complex and of considerable size, reflects the behavior of motor/load combination as determined through time simulations of the same two axis model.

With the algebraic frequency domain relationship, one may explore the sensitivity of the phase currents and the T_{em} to the motor parameters and to parameters which describe the load. Further, one may choose motor and load parameters to enhance the sensitivity of the phase currents and T_{em} to perturbations of the load parameters. It is envisioned that a Rayleigh-Ritz solution procedure may be employed to clearly define

both motor and load parameters for maximum sensitivity of the currents, and thus the detection of variations in load.

It is the variations of these load parameters which are indicative of the health of the driven machine. Thus further exploitation of the method presented herein will quantitatively enhance the evaluation of machine health and the diagnostics and prognostics of motor driven machinery.

References:

1. H. C. Stanley, "An Analysis of the Induction Motor," *AIEE Transactions*, Vol. 57 (Supplement), 1938, pp.751-755

2. P. C. Krause and C. H. Thomas, "Simulation of Symmetrical Induction Machinery," *IEEE Transactions Power Apparatus and Systems*, Vol. 84, November 1965, pp.1038-1053

3. Gerald B. Kliman, Rudolph A. Koegl and Rik W. DeDoncker, "A Demonstration of Sensorless Torque Measurement", *Proceedings of the 49th Meeting of the Mechanical Failure Prevention Group, 1995*, pp. 240-249.

4. Kutzler, Bernhard, *Introduction to DERIVE FOR WINDOWS*, 1996, Kutzler, Austria.

MAIN PROPULSION REDUCTION GEAR CASE STUDY

William A.Watts, Sr. Engineer
DLI Engineering, Inc.
253 Winslow Way West
Bainbridge Island, WA 98110

Abstract: A main propulsion reduction gear in a military ship exhibited no audible or outward signs of a mechanical problem. An Expert System automated diagnostic program was used for routine testing. Vibration spectral data indicated amplitudes at specific frequencies that were low compared to vibration levels for most other machines, but high relative to normal for this case. Routine monitoring with the deployed Expert System provided early detection and diagnosis, and tracked the degradation until corrective action was required. Subsequent inspection of the gearbox revealed a condition that warranted immediate correction to avoid extreme costs. This paper describes the method used to detect the fault, the inspection results, and the benefits derived from such an approach.

Key Words: Case; diagnostic; gears; predictive; vibration

INTRODUCTION: For more than 25 years, the Navy has utilized vibration analysis in planning maintenance for rotating machinery aboard all of the US Navy aircraft carriers. The program makes extensive use of vibration data periodically measured on approximately 400 machines per ship. The surveys are conducted approximately 6 months prior to an overhaul period for quality assurance. Digital data collectors and PC based software are used with an Expert System for automated diagnosis. The ship also uses vibration testing on a periodic basis, usually quarterly, to assess machine condition for on-going condition based maintenance.

In late 1991, a survey aboard one of these ships revealed an intermediate shafting problem based on vibration peaks at one and two times intermediate shaft rotational rate frequency. There was no unusual sound emanating from the gearbox. The amplitudes of the peaks were not very high in absolute terms, approximately the same amplitudes as often measured for the bull gear mesh frequency. However, they greatly exceeded the corresponding baseline values. Also, trending of the vibration data from 1991 to 1996 revealed a degrading condition. The existence of this abnormal vibration over a wide area of locations indicated an intermediate shaft alignment problem rather than a more localized bearing problem. In 1996, an inspection of the gears showed only minor gear tooth wear, however, a subsequent dye check indicated very poor gear tooth contact. Appropriate repairs were made. Subsequent vibration test data showed the unit to be in good condition. By correcting the problem before significant gear tooth wear developed, catastrophic failure and a very costly gear set replacement was avoided.

MACHINE TEST PARAMETERS AND FORCING FREQUENCIES: Vibration measurement and data processing are accomplished using a portable digital data collector

in conjunction with PC based automated diagnostic software. The data collector is activated by scanning a bar code for each measurement location. The data are measured using a triaxial array of accelerometers. Data are sampled and processed within the data collector using an 800-line FFT (Fast Fourier Transform) for each of the three axes in two frequency ranges. Data are downloaded to a PC and then processed automatically. The first step in the process is to order normalize the frequency spectra with respect to the rotating speed of a designated reference shaft. In this case, the order normalized spectra produce frequency scales of 0-5 and 0-100 times Low Pressure Turbine shaft rotational rate. The data are then screened for significant spectral peaks and faults by the logic based Expert System to produce a machine condition report.

Figure 1 illustrates the internal schematic of the gearbox, a locked train double nested gear set. There are four test measurement locations located at mid-height of the reduction gear casing at each of the four corners. The three axes are designated Axial (A), Radial (R) and Tangential (T) relative to the rotating shaft. The measurement locations are designated numerically as 1 and 2 forward and aft on the HP Turbine side and 3 and 4 forward and aft on the LP Turbine side. The data shown later will refer to these numbered locations. Finally, in order to provide for consistent test conditions, a nominal test speed of 110 main propeller shaft RPM has been established.

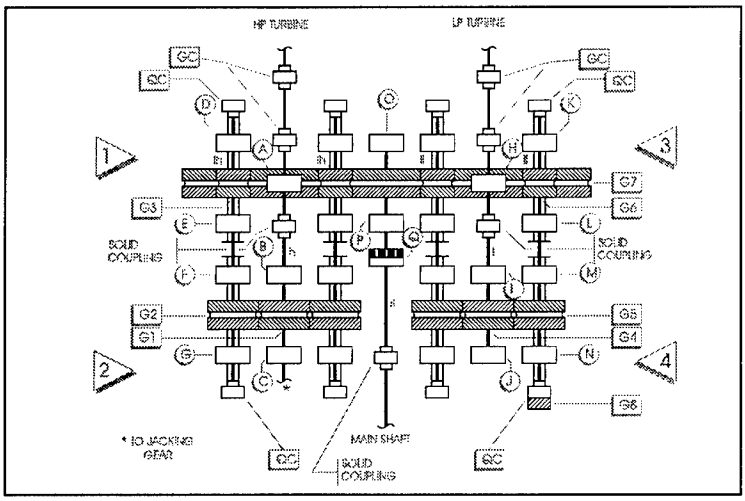


FIGURE 1
VIBRATION TEST & ANALYSIS GUIDE SCHEMATIC

The forcing frequencies generated by this gearbox are easily calculated in terms of the gear tooth counts. For example, the LP Turbine side high speed pinion has 63 teeth driving a gear with 142 teeth on the intermediate shaft. Therefore, the intermediate shaft rotational rate is equal to 63 / 142 = 0.444 times LP Turbine rotational rate. It will appear on an order normalized spectrum at 0.444 orders on the low frequency range graph.

BASELINE DATA: Figure 2 is the baseline spectra for this model of reduction gear for one measurement location (No. 3). It includes the low frequency range (0-5 times LP Turbine rotational rate) spectra for each of the three axes. The high frequency range (0-100 times rotational rate) spectra are not included here. The bottom line on each plot is the computed average amplitude for each spectral line of data. The middle line on each plot is average plus one standard deviation (average plus sigma) and the top line is average plus two times standard deviation (average plus two sigma).

The amplitude scale is logarithmic and is presented in units of Velocity deciBel (VdB). For those of you who think in linear units, 105 VdB is approximately equal to 0.1 inch/second pk. A 6 dB difference is equivalent to a linear factor of 2. Similarly, 12 dB is a factor of 4, 20 dB is a factor of 10 and 26 dB is a factor of 20. For example, 105 VdB (0.1 in/sec) is ten times as high an amplitude as 85 VdB (0.01 in/sec). These relationships will be brought to light for this case later.

The average baseline spectra were constructed from 24 tests for a variety of identical units that were determined via review to be in good condition. The maximum average plus one sigma vibration amplitude for the entire unit is 92 VdB (0.022 in/sec pk) at the bull gear mesh frequency (location 1, radial axis). Thus this gearbox, which is the size of a living room, is inherently a very quiet machine.

VIBRATION ANALYSIS CRITERIA: With vibration analysis in general, it is apparent that the frequency of vibration is usually a good indicator of the source of the vibration. The source then leads to determining which type of mechanical fault is in evidence. The debate more often surrounds the question of what amplitudes are considered to be excessive and the severity of the apparent faults. By establishing average baseline spectra for a specific machine type, we are allowing the machine itself, and many others identical to it, to determine what amplitudes are allowable. Statistically average plus sigma amplitudes represent approximately the 85^{th} percentile. Of the 24 past sets of test data in the average spectra, therefore, one can expect that 20 of the 24 amplitudes at a particular frequency, location and axis were below the average plus one sigma value. As a human analyst and in designing the Expert System, one has to allow for judgment in empirically using this average plus one sigma criterion. If the vibration amplitude at a specific frequency such as a gear mesh exceeds the average plus sigma amplitude, there should not be the automatic assumption that there is a mechanical fault. One looks at the overall picture. By how much does an individual spectral peak exceed the criterion? How many different peaks support the diagnosis and in what pattern ? How high are the absolute amplitudes ? Years of experience reviewing data for various machine types, compounded with a certain amount of actual machine condition feedback, has provided the basis for determining when a fault diagnosis should be revealed and at which points the fault severity should be increased from slight to moderate to serious to extreme. Typically vibration used to support a diagnosis must meet an absolute amplitude requirement and exceed average plus one sigma by at least a few dB to be cited in a diagnostic report even to a slight degree of severity.

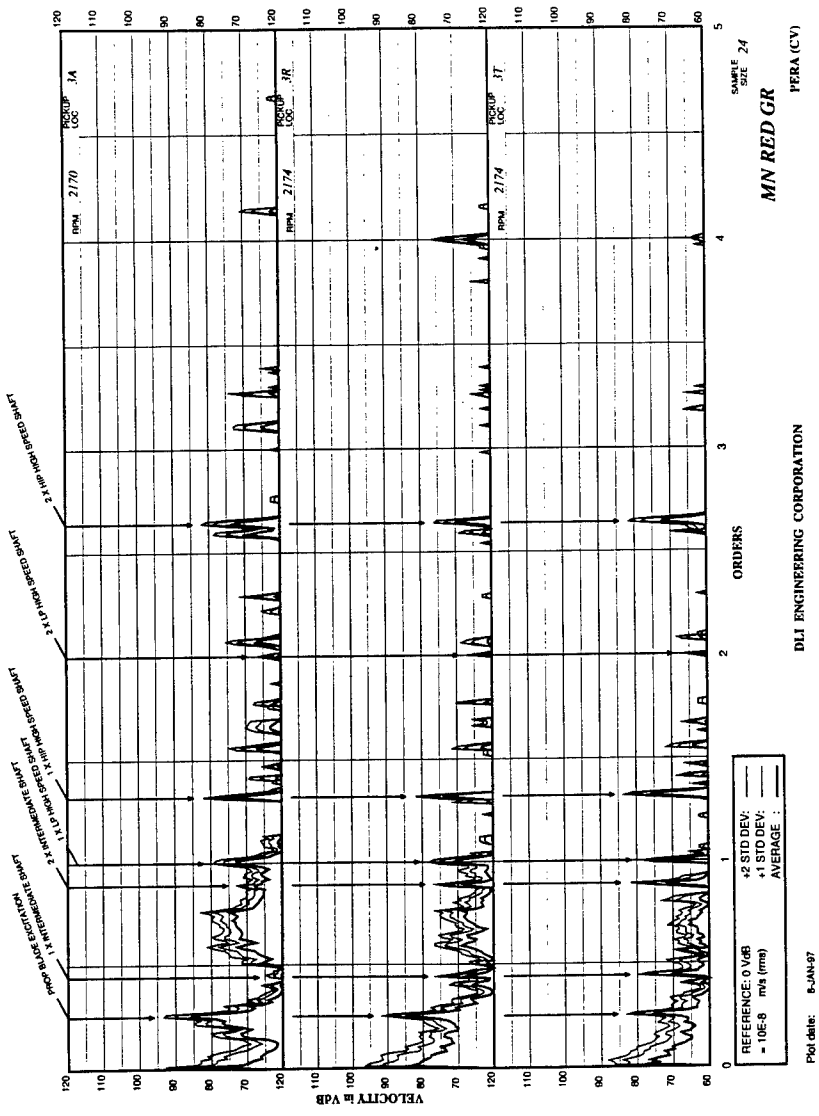


FIGURE 2
AVERAGE SPECTRUM - LOW RANGE

TEST DATA AND RESULTS: Figures 3 is the actual low frequency range data for this unit measured in June 1995. For each spectral plot, the current test spectrum is superimposed on the appropriate average plus one sigma spectrum for this specific machine type. As with the average spectra in Figure 2, the data shown here are only for measurement location No. 3. Note that almost the entire range of each spectrum appears to be quite "normal". Significant exceptions are the spectral peaks at one and especially two times intermediate shaft rotational rate (1xI and 2xI).

Figure 4 is called a screening table and is a tool used for manual analysis and review of test data. Without going into detail, it is sufficient here to state that it is a tabular summary of significant spectral peaks for the given machine test. Included are the vibration amplitudes (in VdB), their dB change from previous, and their variation in dB from average plus one sigma. These values are presented for ten specified frequencies, including 1xI (one times intermediate shaft rotational rate) and 2xI. Note that the 2xI vibration levels throughout the four locations and three axes are generally 20 to 26 dB above average plus one sigma values. This translates to a linear ratio of 10 to 20 times greater than our criteria. Furthermore, the table shows that there was a significant increase in amplitudes from the previous survey. The capability to trend fault severity over time was used to assess the rate of degradation. Similarly, the 1xI levels are generally 12 to 20 dB above average plus one sigma, or 4 to 10 times greater than our criteria.

Figure 5 is the diagnostic report automatically produced by the Expert System used aboard these vessels. The first analysis conclusion drawn is obvious due to the very large exceedances of average plus sigma at 1xI and 2xI. There is a rather serious mechanical problem somehow related to the intermediate shafts. The next observation is that the abnormal amplitudes are fairly equally distributed among the four corners of the gearbox casing. A dominance of the amplitudes and/or exceedances of average at one measurement location would indicate a localized intermediate shaft bearing problem. The equal distribution of the vibration suggests either a major shafting alignment problem or structural resonance. Also, the 2xI frequency (1925 CPM) turned out to be the maximum vibration amplitude measured at the HP Turbine, the LP Turbine and the Attached Lube Oil Pump. A subsequent speed runup test performed by the shipyard showed that the 1xI and 2xI vibration increased with increasing shaft speed, thus ruling out a resonance problem. The final analysis, based solely on vibration spectra, was that there was a serious to extreme intermediate shaft misalignment. The recommendation was to open and inspect the gearbox to check the alignment.

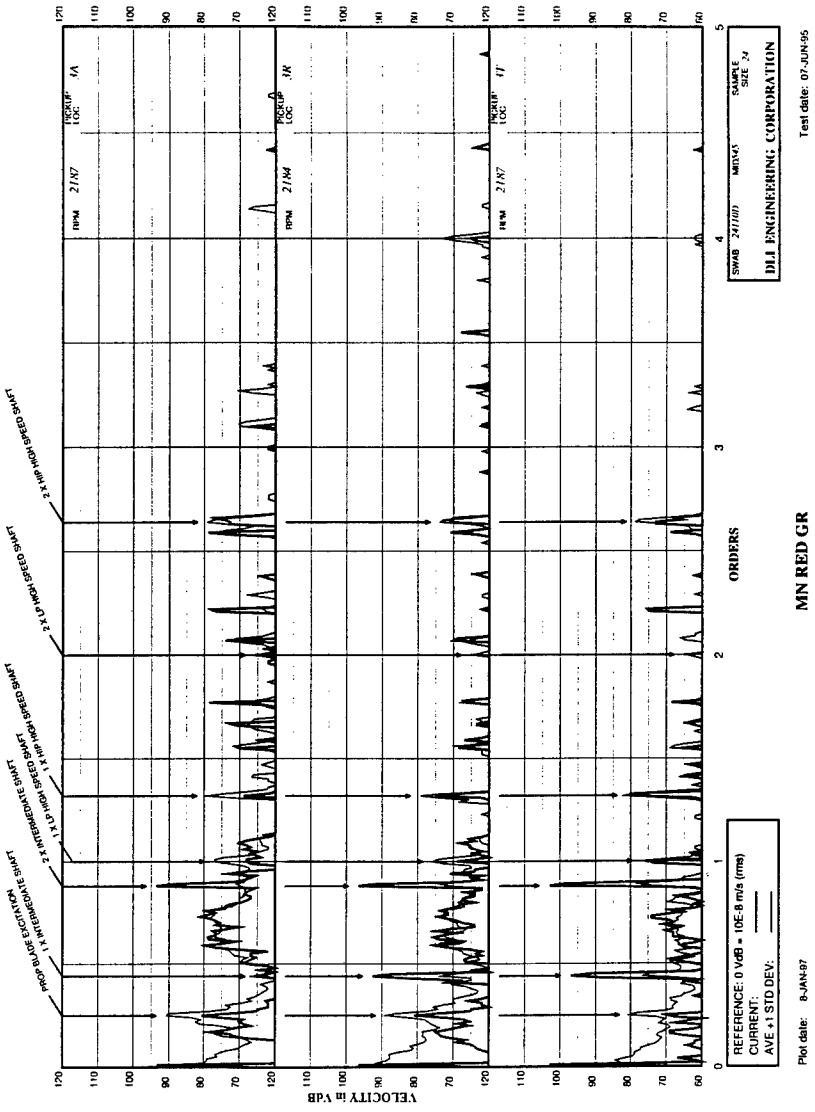


FIGURE 3
TEST SPECTRUM - LOW FREQUENCY

DLI ENGINEERING CORPORATION
MACHINERY VIBRATION ANALYSIS SYSTEM
DATA TABULATION WITH SPECTRAL SCREENING

MN RED GR

PAGE 1
2-JAN-97
MVB010-03

CURRENT SURVEY: 07-JUN-95 ACQUIRED: 06-21-1995 02:23:37 CURRENT RPM: 110 MATRIX ID: SWAB NO: 24110D
PREVIOUS SURVEY: 17-MAY-94 PREVIOUS RPM: 109 SAMPLE SIZE: 24 PLOT RANK: 46

P K P	L V L	SKI SLOPE ALARM	NOISE FLOOR 75%HR	0.25 F3	VDB AT SPECIFIED ORDERS 0.44 1XI	0.89 2XI	1.00 1XL	1.32 1XH	2.00 2XL	2.64 2XH	26.18 LG	63.00 LHG	67.44 HHG	MAXIMUM VDB AT UNSPECIFIED ORDERS LOW RANGE		HIGH RANGE	
1A	CUR	60	48	76	75	86	67	81	62	77	95	50	65	78 0.62	78 0.73	70 7.2	68 20.4
	C-P	-11	0	-2	-8	5	-1	3	3	5	7	2	6	-1	0	-3	-2
	C-A	-10	-12	-15	13	17*	-8	3*	-4	-2	9*	-12	-12	2	1	-3	-8
1R	CUR	63	48	77	95	93	66	70	56	74	96	50	69	78 0.41	76 0.66	75 52.0	73# 8.8
	C-P	-15	0	-4	2	5	4	11	6	7	7	-9	-8	-1	-1	3	1
	C-A	-16	-12	-12	20*	25*	-11	-9	-6	-4	4*	-10	-7	11	3	4*	-5
1T	CUR	60	48	67	96	102	73	81	62	73	83	48	56	74 0.72	73 2.21	68 7.2	68 20.8
	C-P	-5	0	-3	6	9	2	5	6	9	-1	0	0	7	11	-3	-3
	C-A	-14	-12	-13	21*	27*	-4	1*	-2	-7	-6	-12	-12	5	13	-3	-4
2A	CUR	59	48	81	75	84	66	83	64	76	83	50	68	77 0.17	77 2.22	72 21.2	71 8.0
	C-P	-5	0	0	-5	8	-8	7	5	8	0	2	-7	-2	10	-1	0
	C-A	-8	-12	-9	12	13*	-10	4*	-3	-2	-3	-12	-10	-1	15*	-1	-4
2R	CUR	59	50	75	94	89	69	78	56	73	82	62	84	77 1.78	75# 2.22	80# 26.8	75# 12.4
	C-P	-6	0	2	4	14	0	9	-3	2	7	0	4	8	3	-3	4
	C-A	-16	-10	-6	22*	21*	-8	-4	-6	2	-3	-5	1*	14*	13*	-5	-1
2T	CUR	60	50	64	91	101	56	75	48	70	85	48	70	72 1.78	70 2.59	79 12.4	75# 21.2
	C-P	-2	0	-6	4	15	-12	10	-2	5	0	-2	-17	10	5	0	7
	C-A	-14	-10	-14	16*	25*	-23	-4	-14	-1	-7	-15	-1	6	8	3*	-1
3A	CUR	59	48	80	65	93	68	69	65	74	81	56	50	80 0.73	79 0.59	81# 26.8	69 7.6
	C-P	-8	0	-2	-15	1	1	-6	9	4	2	6	0	-5	-1	2	-2
	C-A	-14	-12	-10	2	22*	-9	-10	1	-5	-2	-10	-15	2	1	-2	-3
3R	CUR	64	48	81	92	96	69	72	59	71	91	64	62	75 0.63	74 0.17	72 52.4	72 14.0
	C-P	-7	0	-2	1	8	-1	10	0	4	15	5	14	-4	-5	7	1
	C-A	-16	-12	-8	19*	23*	-6	-7	-6	-2	0	-7	-10	2	-4	7*	4*
3T	CUR	61	48	70	97	103	73	81	56	73	84	48	48	75 2.22	73 0.73	75# 26.8	72 8.0
	C-P	-3	0	-4	4	7	-1	1	-3	9	2	0	0	11	0	-6	1
	C-A	-13	-12	-10	21*	25*	-3	1*	-9	-5	-7	-18	-15	15*	6	-16	4*
4A	CUR	59	48	82	72	94	70	70	64	79#	80	56	48	81 2.21	81 1.77	72 18.0	70# 7.6
	C-P	-2	0	2	-2	5	1	-1	8	6	0	-15	-2	13	10	1	1
	C-A	-10	-12	-9	10	22*	-6	-8	-1	0	-5	-13	-20	21*	14*	-3	-1
4R	CUR	60	48	76	92	98	73	75	59	70	84	76	64	77 2.07	77 2.59	77 14.0	75 16.0
	C-P	-4	-2	-2	4	7	6	3	3	-1	13	7	-1	10	1	-2	2
	C-A	-14	-12	-6	19*	30*	-5	-3	-3	-2	-4	-2	-12	12*	7*	3*	-2
4T	CUR	60	48	68	92	99	56	76	50	71	85	59	56	79 2.22	77 1.77	77 21.2	76 18.0
	C-P	-2	0	-3	3	9	-8	7	0	1	2	-14	0	11	9	-1	-1
	C-A	-11	-12	-12	17*	26*	-23	-3	-10	2	2*	-9	-19	19*	15*	-3	-2
*	xW.F.		0	0	310	552	0	9	0	0	60	0	4	81	77	17	8

TOTAL (* x WEIGHTING FACTOR) x NORMALIZING FACTOR = NORMALIZED FIGURE OF MERIT:1118 x 0.35 = 392 MAXIMUM VDB: 3T 103 AT 0.89X

FIGURE 4
SCREENING TABLE

MN RED GR #
Acquired: 06-21-1995 02:23:37 Speed: 1XL = 2187 RPM (1XS= 111 RPM) Avgs. =24

CLASS D, MANDATORY: CHECK INTERMEDIATE SHAFT FOR IMPROPER ALIGNMENT AND BEARING CLEARANCES

Maximum Level is: 103 (+25) VdB [3T] at 2.00xI

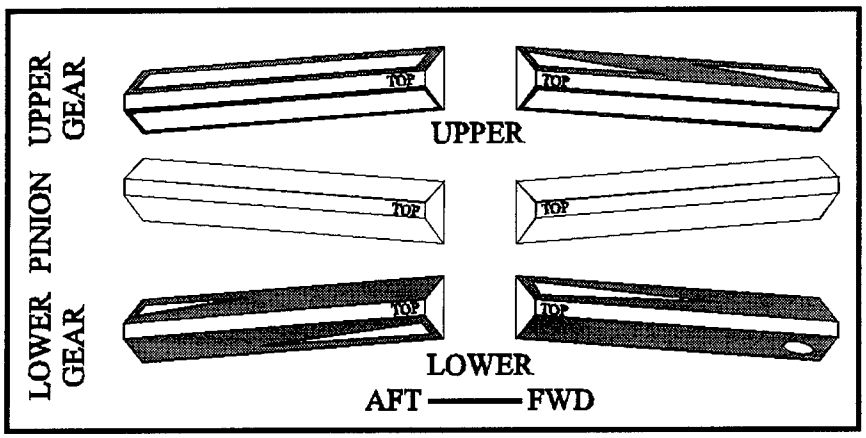
EXTREME INTERMEDIATE SHAFTING PROBLEM is indicated by

95	(+20)	VdB	[1R]	at 1.00xI
96	(+21)	VdB	[1T]	at 1.00xI
86	(+17)	VdB	[1A]	at 2.00xI
93	(+25)	VdB	[1R]	at 2.00xI
102	(+27)	VdB	[1T]	at 2.00xI
94	(+22)	VdB	[2R]	at 1.00xI
91	(+16)	VdB	[2T]	at 1.00xI
84	(+13)	VdB	[2A]	at 2.00xI
89	(+21)	VdB	[2R]	at 2.00xI
101	(+25)	VdB	[2T]	at 2.00xI
92	(+19)	VdB	[3R]	at 1.00xI
97	(+21)	VdB	[3T]	at 1.00xI
93	(+22)	VdB	[3A]	at 2.00xI
96	(+23)	VdB	[3R]	at 2.00xI
103	(+25)	VdB	[3T]	at 2.00xI
92	(+19)	VdB	[4R]	at 1.00xI
92	(+17)	VdB	[4T]	at 1.00xI
94	(+22)	VdB	[4A]	at 2.00xI
98	(+30)	VdB	[4R]	at 2.00xI
99	(+26)	VdB	[4T]	at 2.00xI
81	(+14)	VdB	[4A]	at 4.00xI
81	(+21)	VdB	[4A]	at 5.00xI

FIGURE 5
DIAGNOSTIC REPORT

With the abnormal vibration levels cited in the report, one might guess that the problem would have been noticeable to an observer on site. Keep in mind that the amplitudes cited here are not very high in absolute terms compared to those generated by many other machines in the space, even those that are in good mechanical condition. When standing next to the reduction gear, one is enveloped by a loud symphony of machine noise and the shaking of deck plates. There was no noticeable problem detected merely listening to the reduction gear or feeling the casing.

INSPECTION RESULTS AND SUBSEQUENT REPAIRS: Following the June 1995 vibration test, the gearbox inspection was conducted in two phases. The first phase was an in-port visual inspection of the gear teeth and casing to observe overall condition. The result was that the gear teeth and casing were in relatively good condition, although minor tooth wear was becoming evident. The second phase consisted of a tooth contact check. This procedure involves covering the gears with red dye while in-port. Then the machine is operated under normal loading conditions. A subsequent inspection of the teeth reveals tooth contact patterns using the dye. In some cases, the results showed as little as 5 per cent contact area. Figure 6 is an illustration of the tooth contact pattern. This investigation confirmed the vibration analysis of severe intermediate shaft misalignment.



Areas not shaded represent areas of contact on the gear tooth faces.

FIGURE 6
TOOTH CONTACT

As a result of the vibration test and inspections, the following repairs were accomplished:

1. The high speed flexible couplings between the turbine and reduction gear were disassembled, cleaned, and inspected with respect to dental tooth contact and axial float. The same procedure was performed for the intermediate shaft flexible couplings.
2. All journal bearings on the high speed and intermediate shafts were inspected and scraped as required to improve the tooth contact patterns. Worn bearings were re-babbitted.
3. The gear train timing was adjusted to eliminate the lag of the second reduction pinion.
4. A final tooth contact check was conducted to ensure that the intermediate shaft alignment had been corrected.

POST REPAIR VIBRATION TEST: There was a routine 360 machine MCA (Machinery Condition Analysis) survey conducted following the shipyard repair period. Figure 7 is an excerpt from the screening table for the vibration data measured for this Main Reduction Gear during the survey. Results show that the repairs were quite successful. 1xI and 2xI vibration amplitudes, in particular, all decreased to levels at or below average plus one sigma. There are no mechanical faults indicated.

PKP	LVL	VdB AT SPECIFIED ORDERS	
		O.44 1XI	0.89 2XI
1A	CUR	62	69
	C-P	-13	-17
	C-A	0	0
1R	CUR	79	63
	C-P	-16	-30
	C-A	4	-5
1T	CUR	79	62
	C-P	-17	-40
	C-A	4	-13
2A	CUR	64	71
	C-P	-11	-13
	C-A	1	0
2R	CUR	79	66
	C-P	-15	-23
	C-A	7	-2
2T	CUR	79	70
	C-P	-12	-31
	C-A	4	-6

PKP	LVL	VdB AT SPECIFIED ORDERS	
		O.44 1XI	0.89 2XI
3A	CUR	62	71
	C-P	-3	-22
	C-A	-1	0
3R	CUR	68	67
	C-P	-24	-29
	C-A	-5	-6
3T	CUR	76	65
	C-P	-21	-38
	C-A	0	-13
4A	CUR	59	69
	C-P	-13	-25
	C-A	-3	-3
4R	CUR	65	64
	C-P	-27	-34
	C-A	-8	-4
4T	CUR	75	67
	C-P	-17	-32
	C-A	0	-6

FIGURE 7
POST REPAIR SCREENING TABLE (EXCERPT)

CONCLUSIONS: Some important points and lessons can be learned from this case history:

1. Using narrowband FFT average baseline data for a specific machine type is an excellent criterion for assessing vibration amplitudes and fault severity. General alarm limits or broadband vibrations readings are often insufficient or useless.
2. Using the predictive maintenance philosophy of periodic testing of the machinery allows for the detection of faults when mere observations do not. Periodic testing also allows for trending of faults so that repairs can be planned for a more convenient time.
3. The correcting of mechanical problems before much more costly and irreversible damage occurs will save both money and operations in the future.

OPTIMIZING RELIABILITY, MAINTAINABILITY AND COST FOR SHIPBOARD SYSTEMS

John A. Latimer, William C. Momberger, Nathan Johnson, Jr.
Science Applications International Corporation
3045 Technology Parkway
Orlando, Florida 32826

Abstract: At present, shipboard system reliability is maintained by periodically refurbishing systems at considerable cost per ship. This paper recognizes the need for periodically refurbishing shipboard systems to sustain high reliability, and recommends a Weibull-based analytical procedure for adjusting refurbish interval, and thereby optimizing shipboard system reliability, maintainability and cost, Cosby & Johnson[1]. The analytical techniques offered in this paper exploits the multi-distribution capability of the Weibull distribution, Johnson[2] to analyze actual shipboard system performance data and from these data, determine the actual reliability of the shipboard system, the maintenance cost to sustain that reliability, and the cost associated with the maintenance efforts. To improve existing procedures for on-board equipment maintenance and logistics support, the Navy developed its Maintenance and Materiel Management System (3-M System) during the 1960s, DiStefano[3]. Although the 3-M System provides a maintenance history on all ship system equipment; e.g., main propulsion, electrical generation and distribution and auxiliary machinery, the potential of this database for maintenance cost savings and spare parts allocation have not been exploited to its full potential. The Weibull-based technique that is being proposed for analyzing the 3-M System database is capable of establishing cost effective refurbish intervals, providing the number of shipboard system failures to be expected over refurbish intervals as well as the associated maintenance cost to affect repairs of failed shipboard systems that occur within refurbish intervals.

Key Words: Analyzing; cost; database; Kolmogorov-Smirnov; logistics support; maintenance; shipboard; Weibull; 3-M system.

INTRODUCTION: The U. S Navy experiences logistics problems in their attempt to maintain vital equipment on-board active ships in a high state of readiness. The Navy must provide adequate spares on-board each ship and to this end have developed its Maintenance and Materiel Management System (3-M System) to capture equipment maintenance support data, cost of this support, and provisions for data on which to base improvements in equipment design and spare parts support, DiStefano[3]. This paper proposes to employ a Weibull-based technique to analyze the raw data obtained from the U. S Navy's 3-M System for all shipboard systems and equipment. The technique will be demonstrated by employing a single shipboard system, such as turbines, pumps or valves, to provide a complete estimate of the reliability and maintenance requirements for the life of a particular class of ship. This analysis technique requires that failure time data on each

pertinent shipboard system be extracted from the Maintenance and Materiel Management System. If predictions are to be meaningful this failure time information must be annotated to show that a failure occurred while steaming or in port.

ANALYSIS: Inasmuch as mechanical parts do not in general fail in accordance with exponential theory, the Weibull distribution is employed to analyze the data trend arising from observing 100 sea-valves for a period of 40 months. Within the 40 months, three sea-valve failures were observed: (1) one sea-valve failure after 18 months of service, one sea-valve failure after 27 months of service and one sea-valve failure after 40 months of service, for a total of three sea-valve failures out of a sample of 100 sea-valves (Table I). Weibull analysis of these three sea-valve failures produced a Shape parameter of β = 1.390302 and a characteristic life parameter whose value is η = 476.84 months (Appendix A).

Table I
Observed Shipboard Sea-valve Failures

Usage Interval (Months)	Observed Valve Failures	Cumulative Valve Failures	Predicted Valve Failures
18	1	1	1.051(1)
27	1	2	1.838(2)
40	1	3	3.151(3)

An analysis of these data indicate that the average sea-valve failure rate, measured over 40 service months, is .79714 x 10^{-3}/month with a corresponding Mean-Time-Between-Corrective Maintenance (MTBCM) of 1254.5 months. The Kolmogorov-Smirnov (K-S) confidence bounds ($\pm d_\alpha$) express the probability that the actual shipboard system performance and cost parameter will lie in within the range of values provided (Table II). That is, the probability is equal to "P_α", that the K-S confidence bounds will contain the actual performance value; i.e., failure, and associated cost for the shipboard system assessed. The K-S goodness of fit test is preferred over the traditional Chi-squared (χ^2) goodness of fit test because of its capacity for detecting smaller deviations than the Chi-square test is capable of detecting; in addition, the K-S test is not burdened by the restriction that the raw data be cellularized with at least five data points per cell as is the Chi-square test, Massey[4].

Table II
Spare Valve Spare Parts for the Next 10 Years

Usage Interval (Months)	90% Lower Failure Estimate	Average Failure Estimate	90% Upper Failure Estimate
12	0.0	0.23(00)	12.43(12)
18	0.0	0.60(01)	12.80(13)
24	0.0	1.05(01)	13.25(13)
30	0.0	1.56(02)	13.76(14)
36	0.0	2.11(02)	14.31(14)
42	0.0	2.72(03)	14.92(15)
48	0.0	3.35(03)	15.55(16)
54	0.0	4.03(04)	16.23(16)
60	0.0	4.72(05)	16.92(17)
66	0.0	5.45(05)	17.65(18)
72	0.0	6.20(06)	18.40(18)
78	0.0	6.96(07)	19.16(19)
84	0.0	7.75(08)	19.95(20)
90	0.0	8.56(09)	20.76(20)
96	0.0	9.38(09)	21.58(21)
102	0.0	10.21(10)	22.41(22)
108	0.0	11.06(11)	23.26(23)

Using the performance parameters derived from the first three data points, the predicted spare parts requirement at a ninety percent K-S confidence, Massey[4] for a single ship (housing 100 sea-valves) for the next 108 months (two 54-month refurbish intervals) are listed in Table II.

Table II indicates that for a refurbish interval of five years (60 months) we can expect an average of five (4.72) sea-valve failures and as much as seventeen (16.92) sea-valves to require replacement or servicing within a five year period. Over the life (10 years) of a ship, we expect an average of thirteen (13) sea-valve failures, and as much as twenty-five (25) sea-valves to require replacement over the ten-year period at ninety percent confidence ($\pm d_{.10}$).

SHIPBOARD SEA-VALVE RELIABILITY, MAINTENANCE AND COST: Based on the sea-valve empirical data in Table I, there is a reliability of .97 that the sea-valves will survive a 54-month refurbish interval with schedule inspection intervals every 18 months. Schedule maintenance implies an interval within which an inspection is made and failed sea-valves replaced or refurbished as needed. At the end of 54 months of service, all 100 sea-valves are refurbished or replaced, regardless of sea-valve condition.

If the scheduled maintenance (inspection) interval is doubled, an increase from 18 to 36 months, with the entire sea-valve complement refurbished or replaced after 108 months of service, we can expect a sea-valve reliability of .90.

Assumptions

(A)	Sea-valves Materials	=	$350.00/valve
(B)	Labor Rate	=	$56.25 per hour
(C)	Sea-valve Replacement Rate	=	15 man days

18 Month Inspection Intervals with a Refurbish Interval of 54 Months: With 18 month inspection intervals, we can expect an average of a **four** (4.04) sea-valve replacements (failures) over a schedule sea-valve refurbish interval of 54 months. After each 54-month service period, the entire sea-valve complement is refurbished. The total cost of refurbishing all 100 sea-valves every 54 months of service is determined as follows:

Days required for sea-valve refurbishing	=	104 valves x 15 man days/valve = 1560 man days
Total labor cost	=	($450.00/man day) (1575 man days) = $ 708,750.00
Total cost of sea-valves materials (105) ($350.00)	=	$ 36,750.00
Cost per 54-month refurbish intervals	=	$745,500.00

Over a 9 year service period (108 months,approximately two refurbish intervals), we can expect ten (10) sea-valve corrective maintenance actions/ replacements (failures).

Days required for sea-valve refurbishing	=	210 valves x 15 man days/valve = 3150 man days
Total labor cost	=	($450.00/man day) (3150 man days) = $1,417,500.00
Total cost of sea-valves materials (210) ($350.00)	=	$ 73,500.00
Total cost over 10 service years (108 months) per ship	=	$1,491,000.00

36 Month Inspection Intervals with a Refurbish Interval of 108 Months: If the inspection intervals are increased to 36 months, a sea-valve reliability of .90 [$R_{\text{sea-valve}}$ (T=108 mo.) =. 90] can be sustained. With 36-month inspection intervals, we can expect approximately **eleven** (11.06) sea-valve maintenance actions (failures) over a scheduled

sea-valve refurbish interval of 108 months. The total cost of refurbishing sea-valves every 108 months of service is determined by

Days required for sea-valve refurbishing	=	(111 valves) (15 man days/valve) = 1665 man days
Total labor cost	=	($450.00/man day) (1665 man days) = $749,250.00
Total cost of sea-valves materials (111) ($350.00)	=	$38,850.00
Total cost over 10 service years 108 months) per ship	=	$788,100.00

The cost differential between the 54-month refurbish interval and 108-month refurbish interval is $702,900 per ship. In order to maintain a sea-valve reliability of .97 with a refurbish interval of 54 months, the cost per reliability increment is $100,414.29.

The total cost of maintaining a sea-valve reliability of .97 over a nine year period (with 18 month inspection intervals) for 100 ships is (100 x $1,491,000.00) $149,100,000.00. The total cost to maintain a sea-valve reliability of .90 over a nine year period (with 36-month inspection intervals) for 100 ships is (100 x $788,100.00) $78,810,000. The cost savings realized by extending the refurbish intervals is **$77,319,000.00** for 100 ships.

CONCLUSION: The expected number of sea-valve failures for two refurbish intervals of 54 months (9-year operating period) is 2 x (4.03) = 8.06 (**8**). The expected number of failures for a single refurbish interval of 108 months is 1 x 11.06 (**11**). The 18-month inspection intervals with a 54-month refurbish interval is costing $702,900.00 (approximately $234,300.00 per failed sea-valve) per ship to avoid three additional sea-valve failures over the life of a ship.

RECOMMENDATION: Change the inspection intervals from 18 to **36** months and the sea-valve refurbish interval from 54 to **108** months and realize a cost-savings in sea-valve maintenance of $702,900.00 per ship. These results are obtained without significant sea-valve reliability degradation (R (T = 18 months) = .97 Versus R (T = 36 months) = .90).

References

[1] Cosby, C.S.P. and Nathan Johnson, Jr., "Reliability Apportionment and Cost," SOLE Proceedings,10th International Logistics Symposium, Lake Buena Vista, FL,1975.

[2] Johnson, Nathan Jr., "On Optimizing Maintainability," Microelectronics and Reliability, Proceedings 1977 Canadian SRE Reliability Symposium, Ottawa, Ontario, Canada, 1977.

[3] DiStefano, Renato, "Development of Ship Equipment Maintenance History," 1987 Proceedings, Annual Reliability and Maintainability Symposium, Philadelphia, PA, 1987, p268.

[4] Massey, F. J., "The Kolmogorov-Smirnov Test for Goodness of Fit," Journal of the American Statistical Association, No.42, 1951.

Bibliography

[1] Ellis, Karan E. and Gregory J. Gibson, "Trend Analysis of Repair Times," IEEE Proceedings, Annual Reliability and Maintainability Symposium, Orlando, FL, 1991, p85.

[2] Epstein, Benjamin, "Tests for the Validity of the Assumption that the Underlying Distribution of Life is Exponential, Part I," Technometrics, Vol. 2, No. 1.,1960.

[3] Epstein, Benjamin," Tests for the Validity of the Assumption that the Underlying Distribution of Life is Exponential, Part II," Technometrics, Vol. 2, No. 2.,1960.

[4] Johnson, Nathan Jr. and Cosby, C.S.P., "A General Distributional Approach to Reliability Predictions," SOLE Proceedings, 11th International Logistics Symposium, Philadelphia, 1976.

[5] Weibull, W., "A Statistical distribution function of Wide Applicability," Journal of Applied Mechanics, Vol. 8,1951.

APPENDIX A

SEA-VALVE RELIABILITY: SHIPBOARD

12/16/1996

WEIBULL FAILURE ANALYSIS FOR THE FIRST 3 FAILURE POINTS

WEIBULL PARAMETER ETA = .47684E+03 MO.
WEIBULL PARAMETER BETA = 1.390302

AVERAGE FAILURE RATE = .79714E-03 /MO.; MTBF = .12545E+04 MO.

CORRELATION COEF = .990246; COEF. OF DETERMINATION = .980588

90% KOLMOGOROV-SMIRNOV (K-S) CONFIDENCE BOUNDS/ESTIMATES

** UPPER FAILURE = .11183E-02 /MO. ; LOWER MTBF= .89420E+03 MO.

FAILURE EXPECTATION FOR 120.00 MO. OF SERVICE

USAGE INTERVAL(MO.)	90 % LOWER FAILURE Expectation	AVERAGE FAILURE Expectation	90 % UPPER FAILURE Expectation
12.00	.00	.23	12.43
18.00	.00	.60	12.80
24.00	.00	1.05	13.25
30.00	.00	1.56	13.76
36.00	.00	2.11	14.31
42.00	.00	2.72	14.92
48.00	.00	3.35	15.55
54.00	.00	4.03	16.23
60.00	.00	4.72	16.92
66.00	.00	5.45	17.65
72.00	.00	6.20	18.40
78.00	.00	6.96	19.16
84.00	.00	7.75	19.95
90.00	.00	8.56	20.76
96.00	.00	9.38	21.58
102.00	.00	10.21	22.41
108.00	.00	11.06	23.26
114.00	.00	11.91	24.11
120.00	.58	12.78	24.98

Preemptive Maintenance and Intrinsic Health Monitoring

Donald K. Hoth, Dayton T. Brown, Inc., 555 Church Street, Bohemia, NY 11716

Abstract

In military terminology, a preemptive operation is the action of striking your enemy before he has the chance to strike you. In the support and maintenance of shipboard machinery and equipment, preemptive maintenance will provide ships' personnel with the ability to act on potential machine failure and repair/replacement situations before shutdown and operational crisis can occur. A Preemptive Maintenance system combines the concepts of equipment Intrinsic Health Monitoring (IHM) (a form of condition monitoring) and Predictive Maintenance (PDM) in a manner which produces optimum efficiency in equipment functional operation as well as optimizing logistics in support of operations.

The integration of this kind of an equipment "health" status system within the framework of functional performance will provide a ship's "equipment situation report" with fewer manhours spent in equipment surveillance and will provide greater assurance of continued faultless machinery operation. The great benefit of an IHM system as applied to ship operation is the system console presentation that continued "fail-free" operation can be expected over a future demand period. (The assurance of planned operations without interruption.) In addition, remote observation of impending problems at a central maintenance support facility can assist in the diagnosis of the problem and the most efficient corrective action.

Introduction

We assemble equipment into functional sets in order to accomplish specific tasks. The design of these equipment sets supports or creates a process which implements an operational need. Our first concern is correct operation. Our second concern is process protection to ensure that operation will continue as needed. Equipment maintenance is our method of process protection and our method of implementation of maintenance determines our state of readiness and a major portion of our operation and support costs.

Maintenance Concepts

Predictive maintenance the newest concept has grown rapidly because of compelling economic factors. Traditionally, the maintenance of industrial equipment was performed in one of two ways: run-to-failure, or preventive maintenance. The run-to-failure approach can result in accidents (poor safety) as well as producing long costly equipment down time. It is also totally undesirable in situations where process interruption leads to coincidental damage or prolonged performance curtailment. Preventive maintenance requires an equipment supervisor or process manager to make some conservative assumptions as to when failure is likely to occur and then to maintain the equipment and provide replacement accordingly. Since machines in predictve and preventive programs are actually maintained, repaired, or replaced before such actions are really necessary, this

type of maintenance can be costly and still fail to catch latent (factory errors) or induced (installation misalignments) failures.

Predictive maintenance presently relies on expert knowledge of specific equipments (derived through testing and historical data) and periodic inspections which include oil samples and vibration analysis/monitoring in order to track equipment deterioration. Vibration analysis/monitoring has been performed with portable data collectors (PDCs). Analysis is performed either within the device or in post-time at a central processing computer. These hand held PDC devices contain a probe which is placed against a machine and held there for a few seconds to measure vibration amplitudes. The frequency of monitoring may be daily, monthly, or even on a quarter yearly basis. PDCs have several drawbacks. They are labor intensive, their diagnostic capability is limited and they capture only a brief snapshot of an equipment signature. (PDC do not usually take into account variable loading conditions.) Environmental conditions, such as temperature or physical accessibility can affect both the frequency and accuracy of the readings. Finally, their accuracy is dependent on the skill level of the individual technician.

Preemptive Maintenance

A typical Preemptive Maintenance system consists of a remote central operation site and a number of monitored facilities. The system will operate as shown in the functional diagram of Figure 1. The central site computer, video display and data printer serve as the IHMS central data collection equipment. Additional software programs are used for maintenance/repair decisions, failure cause analysis and report generation. The central operation site is also used to store the historical data generated by the Intrinsic Health Monitoring System, the diagnostic predictive maintenance data and the preventive maintenance schedules for each local equipment site in the system. Thus, the central site operational staff can observe the equipment status of an unlimited number of remotely located facilities, command centralized assets and coordinate the maintenance and operational activities in an efficient manner. The key to this ability is the IHMS telephone communications and command capability. This communication link allows a great degree of management participation and flexibility in operational process protection.

The system approach of Preemptive Maintenance is to use the IHMS data as an early warning system. Various predictive techniques can then be employed to determine the cause of the impending problem and identification of best corrective action can be determined by a data search in the central workstation computer through equipment historical data/available assets log. Directed maintenance can then be dispatched on a scheduled and prioritized basis.

IHM Concept

Operating equipment and entire machine processes (larger configurations of operating equipment) generate a set of intrinsic physical signatures which are products of, and subsidiary to, their primary performance characteristics. These signatures are nonfunctional parametrics which are consistent and reliable indicators of "normal" operation. Careful perception and analysis of these physical parameters can confirm

equipment and/or process operational status (normal/not normal) and neutral processing of the parametric data can provide indications of future failure or impending out-of-tolerance performance of the operational equipment. (Small perceptible changes in vibration, mechanical stress, or temperature levels will be precursors to failure events.)

IHM System Operation - Monitor Mode

An IHM system effectively combines the IHM concept (for determining whether the observed equipment is operating within its "norms") and the necessary remote data gathering capability required to perform PDM when the equipment is "out of norm". The typical IHM system would be configured as shown in Figure 2. It consists of a central workstation and variable number of IHMU local stations. The IHM units (IHMU) are the primary data gathering elements in the system. Each IHMU gathers data via its sensor complement and transmits this data to the system central workstation. The data is gathered periodically (typically 10-15 minute intervals). The IHMU also processes the periodic data readings (PDR) and performs an appraisal of the equipment status (normal/not normal) and transmits this status or condition evaluation at hourly intervals. These hourly condition evaluations (HCE) are formed in the IHMU which is an Intelligent, P.C. based, local processor and no further processing is required at the central workstation. The central workstation maintains two data bases: one for the equipment status displays (the HCE data); the second is the PDR (for analysis, when necessary, by expert knowledge predictive maintenance techniques).

As shown in Figure 3, the operating subsystem equipment chosen for display is a typical conventional steam turbine propulsion plant. Figure 4 is the typical subsystem status display which shows the condition/status (in color) and the IHMU analysis of future condition (in the form of a probability of failure or a failure potential index [FPI]).

The central workstation can also save the "normal" data sets and specific failure patterns in order to build a library on particular equipments for comparison with future data from other similar/identical equipments. This comparison capability allows the system operator to detect new equipment (repairs/replacements) misalignments or improper installations and to possibly detect/diagnose the recurrence of previously noted failures.

Figure 5 is a typical IHM FPI printout produced during a test run in which a motor-pump was induced to fail (the pump impeller was caused to erode by the introduction of a caustic element into the water flow). As the readings show, it took 5 days for the pump to fail (loss of pressure and flow which could not be restored by valve control or motor effort). The IHMS detected the probable failure 168 hours prior to failure and indicated definite future failure 120 hours before the loss of operational performance.

In the past two (2) years IHMS has produced 7 to 19 day warnings in six (6) different industrial trials which encompassed 40,000 hours of operation.

IHM System Operation - Learn Mode

In order for the IHMS to monitor norms and detect significant deviations from normal, an initial "learning" period is performed. The IHMU takes data on the operating equipment and forms statistical norms for each sensor and for several frequency bands of vibration.

(The exact number of bands is variable/setable for each IHMU.) The data intervals are identical to the periodic data intervals in monitor mode. During learn mode, the only data transmitted to the central workstation is a Normal Condition Confidence Level (NCCL) which is calculated by the IHMU and is a figure of merit on the degree of stability of the PDR data being observed. If, after a 6-12 hour learning period the NCCL indicates stability, the IHMU is placed in monitor mode and the preemptive maintenance operation proceeds. If the NCCL indicates that the learn mode data is not sufficiently stable, the choice and location of the system sensors must be checked and the potential of a preexisting fault investigated.

IHMU Design

The IHMU incorporates a modular open architecture which allows the unit to be configured with a variable set of sensors and sensor processor capability. Typically, an IHMU will consist of a set of specific function printed circuit boards which provide programmable functions such as band pass filtering, programmable gain amplification, sensor signal conditioning and A/D conversion of sensor data. A typical IHMU configuration is shown in Figure 6. The IHMU software, as previously described, has two operating modes and the following detail characteristics:

Monitor Mode

In normal mode operations, the operating software will operate as a "ring" commutator. It will provide continuous data collection from the sensor analog/digital converter channels on a cyclical basis and transmit this data to the neural processors within the cycle. The collect and transmit cycles will be repetitive, timed, and controlled. The normal mode operating software will also collect trend data from the neural processors and store this data in the IHMU memory. As data is collected from the neural processors, it will be examined and an equipment "health" status will be determined. Significant event occurrences and averaged health status data will be stored in the system's status display buffer registers in order to allow operator observation of system trends and events.

Learning Mode

When the system is placed in the learning mode, the operating software will perform in an identical manner to normal mode operations, except that the software will not store or examine the neural processor outputs. The system will examine vibration sensor bandwidth data and calculate the frequency bands of significance. The significant frequencies and energy content will be stored for normal mode operation. The temperature and operating load sensors will be used to adjust or compensate the vibration readings.

Communication Periods

The operating software communication will be controlled by two system counters: the sensor scan counter and the data collect counter. The sensor scan counter (I counter) will control the rate/time period of the sensor scan cycle. The data collect counter (D counter) will control the rate/time period of the neural data collection and storage. Both counters are software adjustable. The operating system software flow chart is shown on Figure 7.

Conclusions

A preemptive maintenance program will provide shipboard operations with a better system for assuring continued operational performance and a better appraisal of planned operating period capability. Ship and shore maintenance personnel will be capable of predicting O & S requirements on a "need" basis rather than on a costly preventive basis. The specific operational benefits will be:

a. Early warning of impending equipment breakdown or out-of-tolerance performance (operations protection).
b. A real-time, local picture of equipment "norms" in regard to equipment maintenance decisions.
c. Equipment/operational trend analysis (long term; short term) for optimal equipment life and reduced system stress/system failure.
d. Remote real-time observation of local equipment conditions.
e. Early disclosure of poor installation, poor repair or defects in newly installed equipment (comparison of historical data for similar equipment).

Specific benefits to maintenance and support will be:

a. Reduced O & S, based on condition-based predictive maintenance and overhaul.
b. Standardized record-keeping with quality data, and easy accessibility.

The need to streamline organizations raises the need for the generation of maintenance management software where machinery condition data is stored and repair/replace/reinspection decisions are evaluated. Installation of local- and wide-area networks further supports this trend towards information sharing. The total benefit of preemptive maintenance is in *improved anticipation of maintenance needs* and *greater confidence in operational status* (present and future).

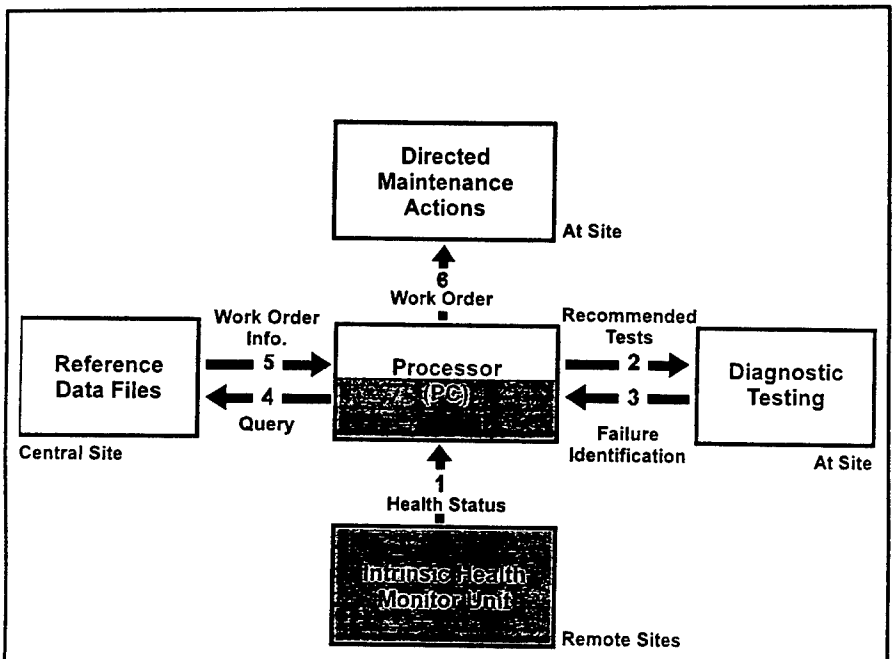


Figure 1 - Preemptive Maintenance Architecture

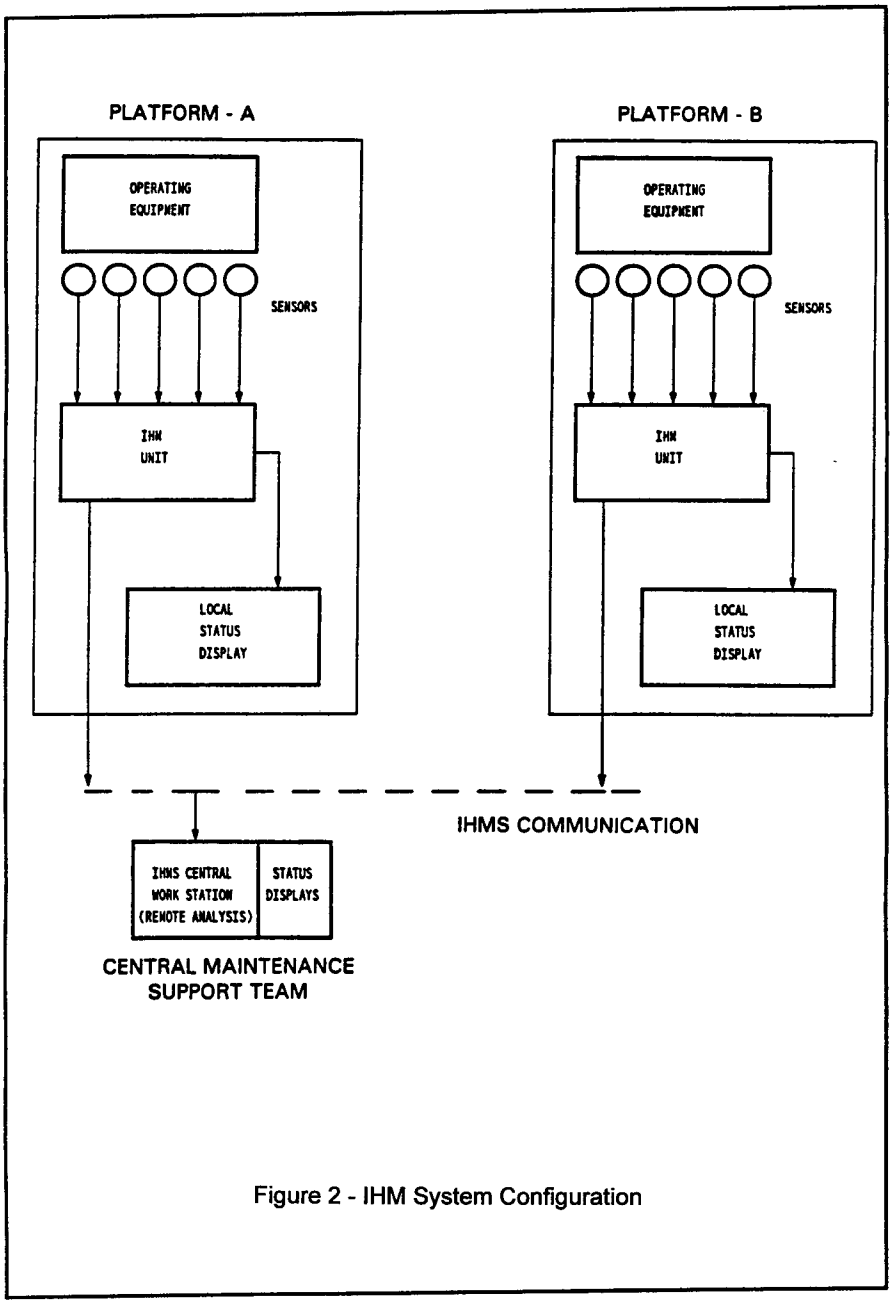


Figure 2 - IHM System Configuration

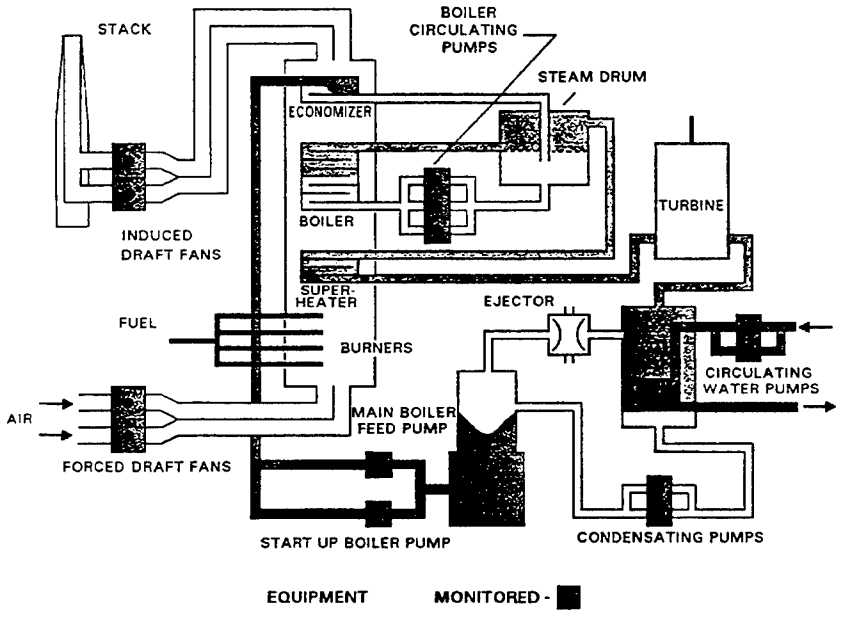


Figure 3 - Typical Application Display

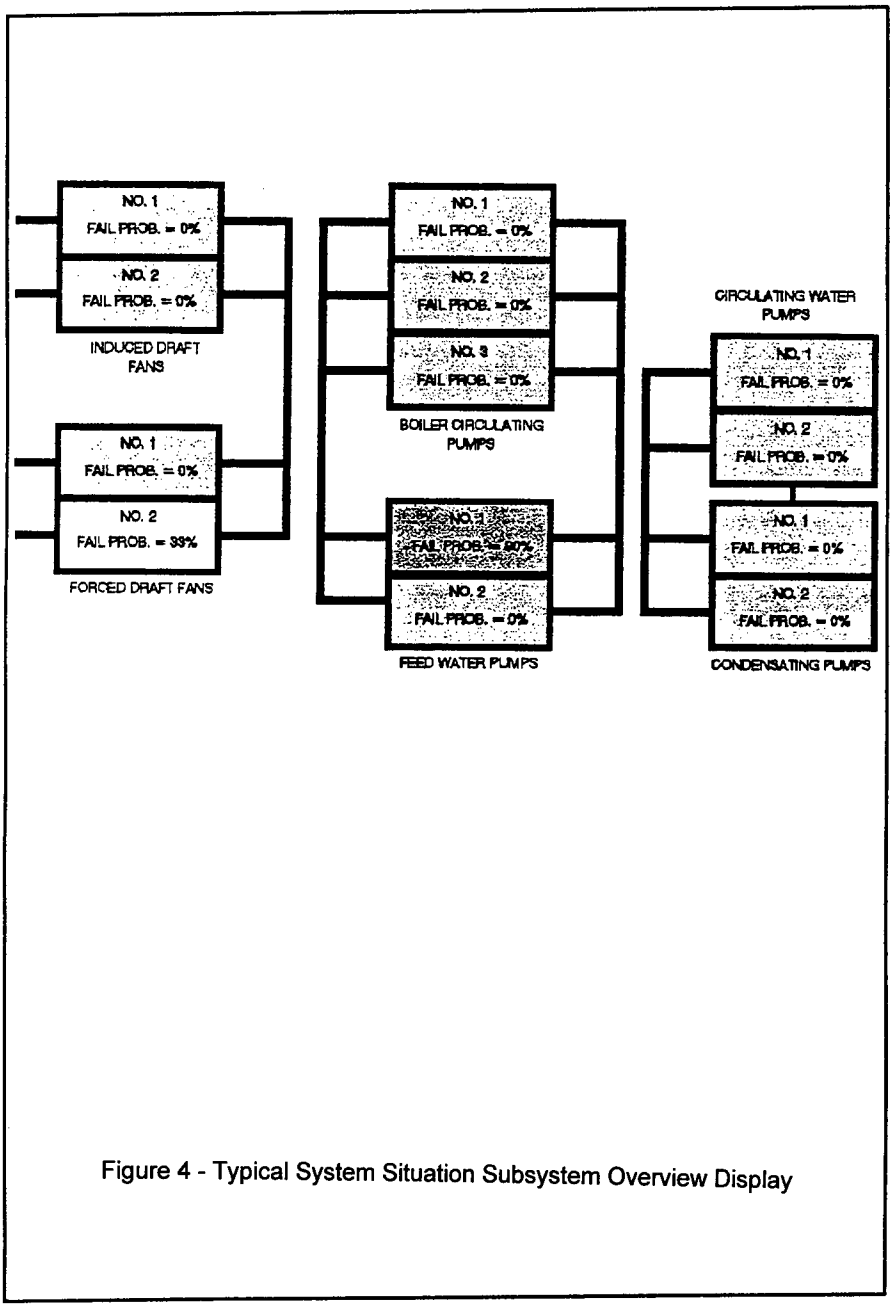


Figure 4 - Typical System Situation Subsystem Overview Display

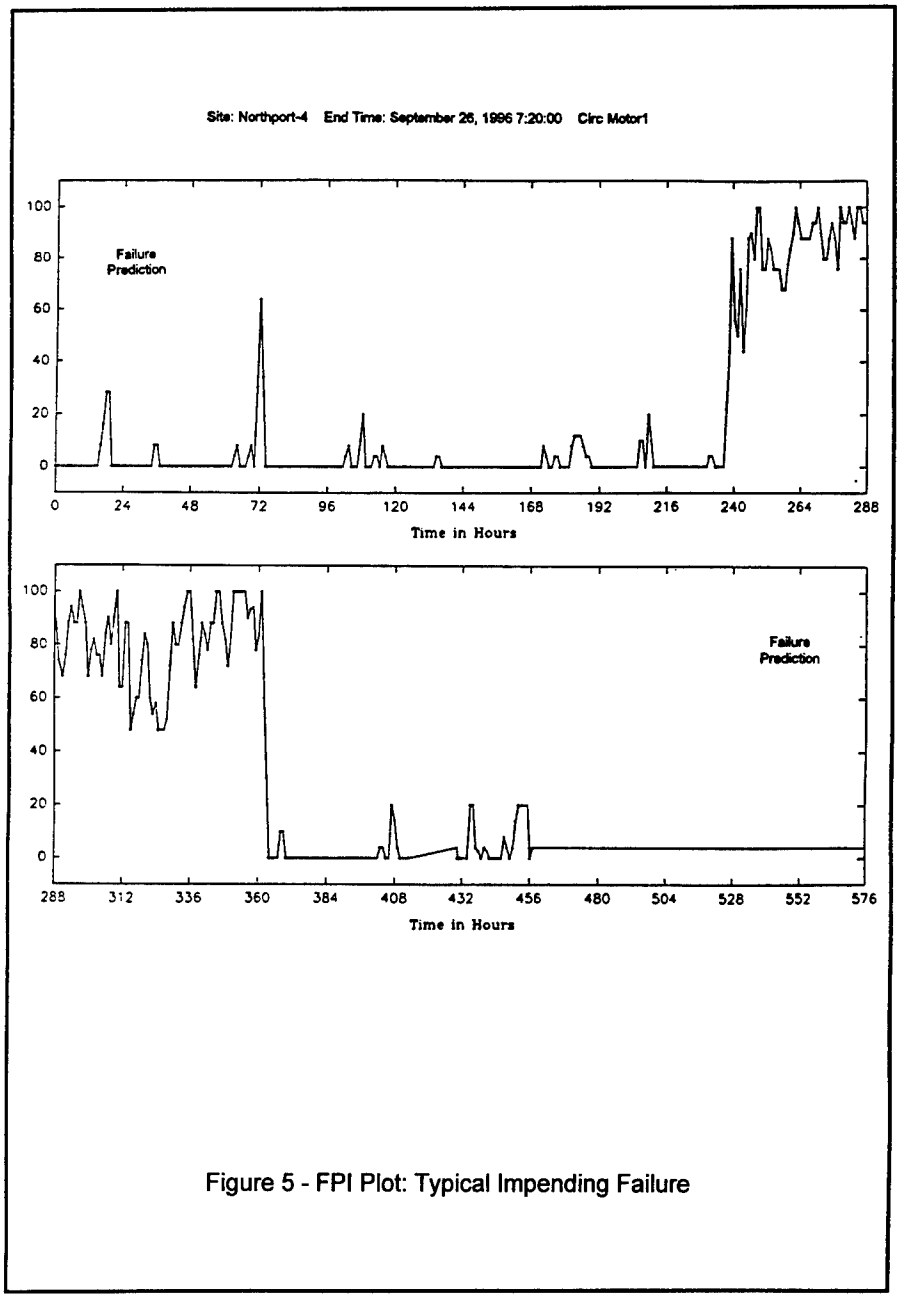


Figure 5 - FPI Plot: Typical Impending Failure

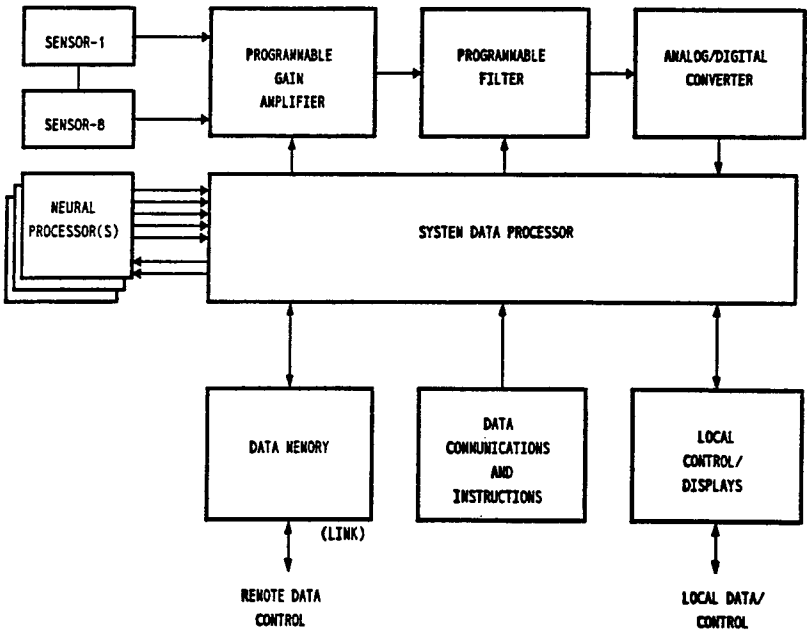


Figure 6 - IHMU Configuration

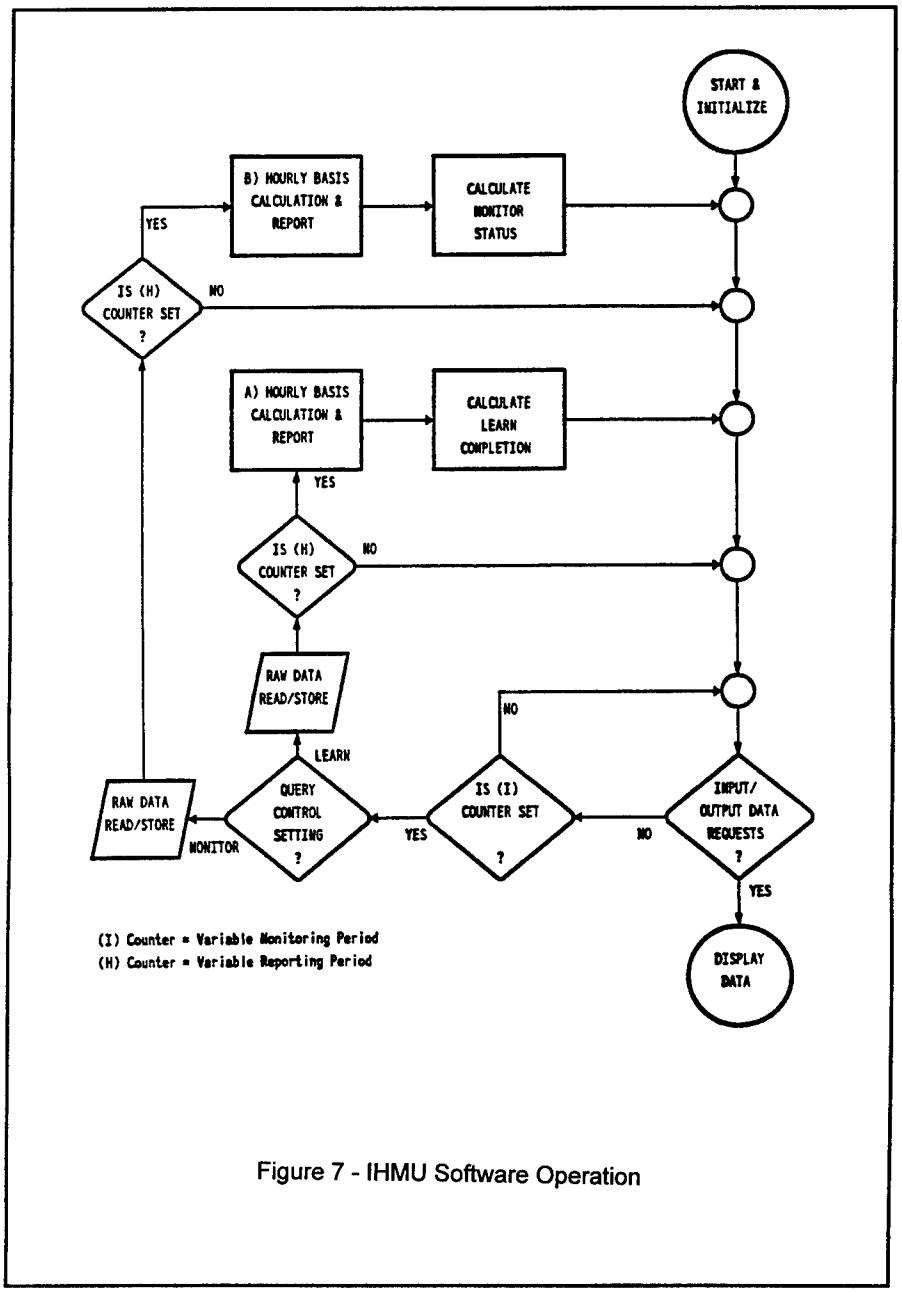


Figure 7 - IHMU Software Operation

RSAFP

PROBABILISTIC DESIGN, FAILURE ANALYSIS & SAFETY METHODS

CoChairs: **B. C. R. Ewan**
University of Sheffield

M. Moatamedi
University of Sheffield

A G-FUNCTION-BASED PROBABILISTIC DESIGN METHODOLOGY FOR SYSTEM RELIABILITY

Wei Wang and Y.-T. Wu

Probabilistic Mechanics and Reliability Section
Southwest Research Institute
P.O. Drawer 28510
San Antonio, TX 78228-0510

Abstract: A new method is developed in this paper to design structures with system reliability constraints. In this method, the g-functions instead of the reliability functions are approximated around the current design and their corresponding MPPs (most probable points) with respect to both the design and random variables. Design variables that are parameters of a random variable are separated from the random variable through a simple transformation. The probability of failure of a structural system is computed using an efficient importance sampling technique that approximates each limit states using a quadratic polynomial with adaptive curvature and position. This system reliability analysis method provides both the accuracy and efficiency needed for the purpose of probabilistic design. A cantilever beam subjected to lateral and vertical random loading is designed using the conventional and the proposed method against excessive lateral and vertical deflection and yielding. The proposed method is five times more efficient than the conventional approach. At the optimal design not all the failure modes are equally important. Therefore, it is important to consider the multiple failure modes as a system not individually during design.

INTRODUCTION: Reliability-based design has the potential of striking an optimal balance between the requirements for reliability and economics, as such it is superior to deterministic design which cannot properly treat the inherent uncertainties involved in every aspect of the life span of a structure, from design to demolition. Yet, reliability-based design still remains in the limbo of research and development after many years of study. One of the main obstacles that is preventing reliability-based design methodology from being widely used in engineering design practice is its consumption on computing resources.

Before we introduce new and more efficient methods, let us first review the conventional way of solving reliability-based design problems. A probabilistic design problem can be formulated in a variety of ways depending on what are treated as objectives, what as constraints and if there are more than one constraints. A popular formula is given below:

$$\begin{aligned} &\text{Minimize:} && \textit{Total Cost} \\ &\text{Sub. to:} && P_f = \Pr[\bigcup g_i \le 0] \le \text{limit} \end{aligned} \tag{1}$$

where g_i's are the limit state functions and define failure when $g_i < 0$ and safety when $g_i > 0$. P_f is the structural probability of failure, and the reliability of the structure, $R = 1 - P_f$.

Total cost is the sum of initial design cost, construction cost, maintenance cost, failure cost, etc.

P_f can be computed with a number of reliability analysis methods, which all involve repeated calculations of the response functions (g). For any given design, **D**, P_f assumes a single value. Therefore, P_f is just a regular deterministic function of **D**. Consequently, Problem (1) can be solved using any standard deterministic optimization algorithms such as SLP (sequential linear programming) [1]. We will call this approach the conventional approach. Since most non-Monte Carlo methods requires the location of the most probable point (MPP) or design point on a limit state surface, reliability analysis is of iterative nature [2]. Consequently, the conventional reliability-based design methods require nested loops, the inner loop computing the reliability and the outer loop updating the design. When the computation of a response function involves complex finite element models, as it often does in the real world structural engineering problems, the demand on computer time to accomplish reliability-based design optimization using conventional approach can be forbidding. This problem becomes even more pronounced when more the one correlated failure modes are involved.

As a result, how to improve the efficiency of probabilistic design optimization has been a constant focus of research [3-6]. Almost all the promising new methods involve using approximate reliability function ($R = 1 - P_f$). There are basically two ways to approximate the reliability function. The conventional way is to approximate the reliability function directly, e.g., linear expansion of P_f with respect to design variables. The problem with this approach is that it is very costly to obtain the expansion, since every perturbation to a design variable demands a full reliability computation. An alternative is to approximate the g-functions with simpler closed form expressions, and then compute P_f using the simplified g-functions. This approach falls into the category of approximating a response function (P_f) through the use of intermediate response functions (g).

In most real world structures, more than one failure modes are present. The design usually requires that the system reliability be above a certain level. Because of the difficulties involved in system reliability evaluation, for many years, system reliability constraints in a design optimization problem are converted to single mode constraints by (1) treating each mode individually through assigning an common, or rather arbitrary, limit on the occurrence of each mode [3,7], and (2) using a simple first-order bounds on system reliability that essentially neglects the correlation among the failure modes [4]. With the availability of more efficient system reliability analysis methods and more powerful computers, such simplification is no longer a necessity.

In this paper, a new reliability-based design methodology that uses adaptive approximate g-functions developed by Wang et al., [5] is extended to design optimization problems with system reliability requirement. The structural system reliability is computed using a curvature based importance sample technique [8] that can provide accurate estimation of system reliability efficiently. A simple cantilever beam subjected to both lateral and

vertical random loading is designed against excessive deflections and yielding using the proposed and the conventional approach. Since by far the most time consuming part in a probabilistic design optimization process is structural re-analyses for most structural engineering problems, the number of structural analyses necessary to complete the optimization is used as a criterion to compare the efficiency of the proposed approach and the conventional approach.

THE DESIGN METHOD: The procedure of the proposed approach is given in Fig. 1. Some of the distinct characters of the proposed method are:

(1) the g-functions rather than the reliability function ($R = 1 - P_f$) are approximated,
(2) the g-functions are expanded with respect to both the design variables and the random variables,
(3) the g-functions are expanded not about the mean values but about the current MPPs that changes from iteration to iteration, and
(4) the random variables and design variables are separated. For example, if the mean and variance of a random variable, X, are design variables, a new random variable, Y, with zero mean and unity variance can be defined by the well know transformation

$$Y = \frac{X - \mu}{\sigma} \tag{2}$$

where μ and σ are the mean and standard deviation of X, respectively

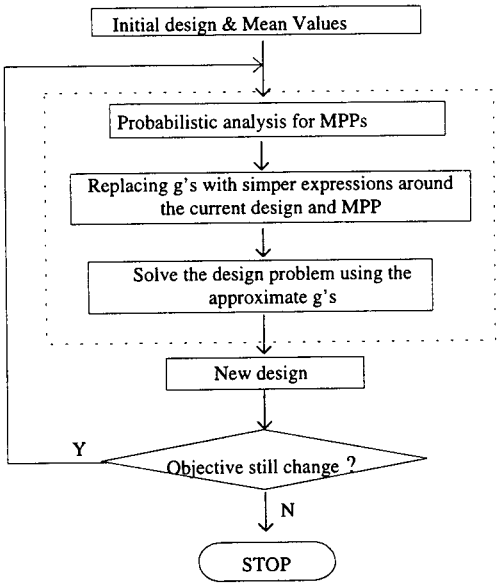


Figure 1. Procedure of the proposed method

Since the approximate g-functions are explicit functions of the random variables and the design variables, the system probability of failure can be computed with these approximate g's during a design iteration without having to performing exact structural re-analysis. As a result, structural reliability computation using the exact g-functions needs to be done only once per design iteration to obtain the current MPP in the proposed method. In contract, during each iteration in the conventional approach, reliability needs to be computed many times using the exact g-functions either to obtained the approximate reliability function or to perform line search. For methods that expand g-functions with respect to only the random variables, new expansions are needed whenever a trial design is attempted.

Many different ways exist to approximate the *g*-functions. The one that linearizes the *g*-functions about the current design and MPP is termed SLg (Sequential Linearization of g-function) [5]. Higher order approximation using information from the previous design points may improve converge rate in some cases, but SLg will be used in this study.

SYSTEM RELIABILITY EVALUATION: There are many methods for system reliability evaluation. The authors believe that methods that mixes the analytical determination of the MPPs and the importance sampling concepts hold the promise to deliver both the needed efficiency and accuracy for reliability-based design. One such method, termed curvature-based adaptive importance sampling (AIS2), is developed by Wu [8] and implemented into the NESSUS probabilistic analysis code developed at the Southwest Research Institute. In this method, the MPP of every failure mode is first located through an analytical search procedure in the u-space (the space spanned by independent standard normal random variables). A quadratic polynomial is formulated around each of the MPP to approximate the exact failure surface, without using the second order derivatives. Samples are then taken around the MPPs. It is possible that the failure region defined by these quadratic polynomials do not cover the entire failure region. The curvature and position of these polynomials are therefore adjusted during the computation using information from the previous samples to make sure that the entire failure region is covered. The concept of this method is illustrated in Fig. 2. This method is used in this study to evaluate the system probability of failure.

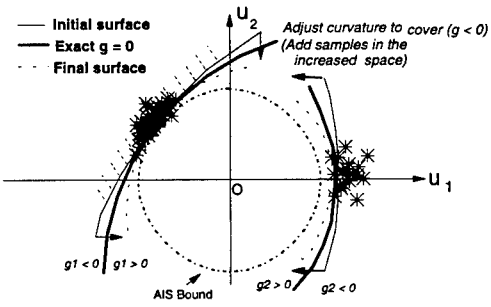
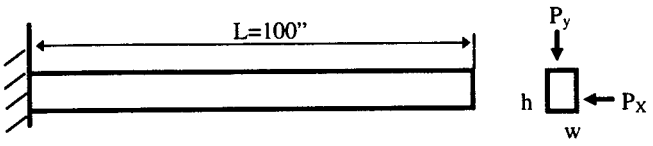


Figure 2. A curvature-based importance sampling technique for system reliability estimation

DESIGN OF A CANTILEVER BEAM: Consider a cantilever beam subjected to both lateral and vertical loading (Fig. 3).



Random Variables (units: psi):

P_X – N(500, 100) P_y – N(1000, 100)

σ_p – LN(40000, 2000) E – LN($3x10^7$, $1.5x10^6$)

Figure 3. Beam Under Vertical and Lateral Bending

Three possible failure modes are considered here, failure due to excessive lateral and vertical deflections and due to yielding. The limit state functions are, respectively,

$$g_1 = 0.5 - \frac{4P_x L^3}{Ew^3 h} \tag{3}$$

$$g_2 = 0.5 - \frac{4P_y L^3}{Ewh^3} \tag{4}$$

$$g_3 = R - (\frac{6L}{w^2 h} P_x + \frac{6L}{wh^2} P_y) \tag{5}$$

where the 0.5 is the limit for lateral and vertical deflection in inches.

The goal of the design is to reduce the weight/cross sectional area of the beam while keeping the probability of failure below 0.00135 (reliability index above 3). This design requirement can be expressed in terms of the mathematical programming language as:

$$\begin{aligned} &\textit{Minimize:} && Obj = wh \\ &\textit{S.t.:} && P[g_1 \le 0 \cup g_2 \le 0 \cup g_3 \le 0] \le 0.00135 \\ & && 2 \le w \le 6 \quad 2 \le h \le 6 \end{aligned} \tag{6}$$

The problem is first solved using the proposed method. The deterministic minimization using the approximate g-functions is carried out using the popular DOT (Design

Optimization Tools) program [9] with the SLP algorithm. Design iteration starts at a conservative design (w = 5, h = 5). The optimum design obtained is (w* = 3.70", h* = 4.62") with a corresponding cross sectional area 17.07 in^2. To check whether the result is the true optimum, a contour plot for the probability of failure and the objective is drawn (Fig. 4), which confirms that the algorithm indeed yields the true optimum. The minor wiggles in the probability contour is because the number of samples used in evaluating system reliability is small.

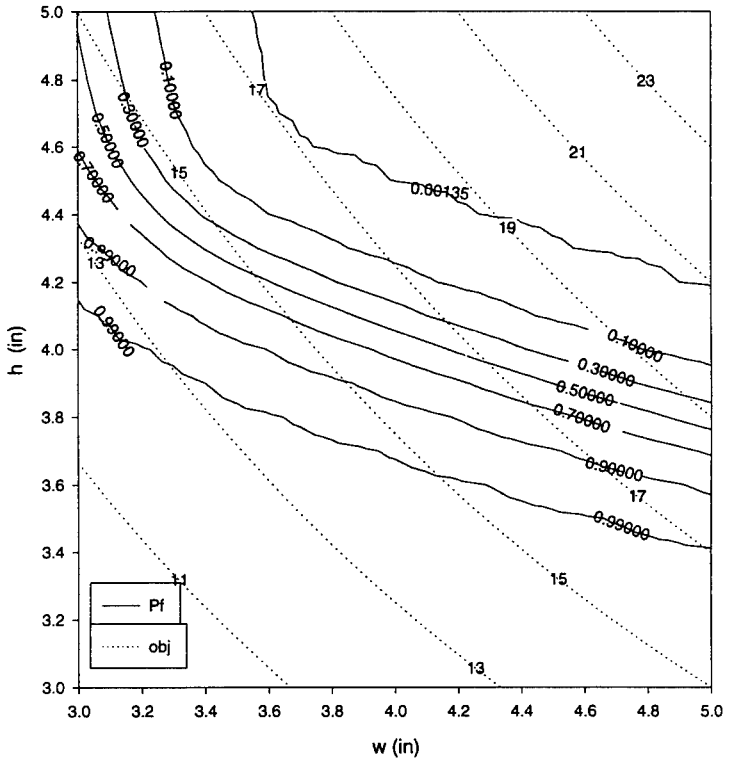


Figure 4. Contours for P_f and Objective

The same problem is solved using the conventional approach, i.e., treating the reliability constraints as a regular constraint in a deterministic optimization problem. The SLP technique in the DOT program is used to carried out the minimization process. It takes 5 iterations and a total of 31 *reliability* computations for the iterations to converge to the same results as obtained by the proposed method in 3 iteration.

To compare the efficiency of the two methods, we need to know the number of exact g-function calculations for each reliability analysis. It is difficult to predict accurately how many exact g-function calculations are needed for each reliability analysis because the number of iterations needed to locate the MPPs and the number of samples needed to estimate the system reliability may vary from design to design. To give a lower bound estimate for this number, let us make the following assumptions:

(1) No exact g-function calculation is needed once the MPPs are obtained, which is true only when the failure region defined by the initial quadratic approximate g-functions covers the entire true failure region.

(2) In average, 2 iterations are needed to locate the MPP for a failure mode using a first order approach (FORM, AMV+, etc.)

Then, the number of g-function calculations for each reliability analysis is 2 (iterations) x 3 (failure modes) x [4 (random variables) + 1] = 30. The number, 30, is a very conservative estimate, i.e., in general, more g-function calculations are needed for one reliability analysis.

If we use this rough estimate, the total number of structural analysis needed to accomplish the design using the conventional approach is 31 x 30 = 930. In contrast, it takes 3 iterations for the proposed method to converge. In each iteration, one reliability analysis to obtain the MPPs and 7 x 3 exact g-function calculations to expand the g's are needed, adding up to a total of 3 x (30 + 3 x 7) = 151, which is 5 times more efficient than the conventional approach. Of course, the efficiency of the conventional method may be different if a different computer program and minimization techniques are used. However, DOT program is a good optimization code and the SLP method is among the most efficient optimization techniques. Therefore, the authors are confident that the efficiency of the conventional method obtain here using DOT and SLP is representative.

The beam is also designed by considering each individual failure mode alone. The results are listed in Table I.

Table I: Design considering different limit state functions

Modes	*g1 only*	*g2 only*	*g3 only*	*g1 U g2*	*g1 U g2 U g3*
w	4.80	2.00	2.72	3.70	3.70
h	2.00	5.65	3.51	4.62	4.62
Obj	9.60	11.30	9.55	17.07	17.07
P_f	0.00135	0.00135	0.00135	P_f= 0.00135 P_{f1}=0.000286 P_{f2}=0.00111	P_f= 0.00135 P_{f1}=0.000286 P_{f2}=0.00111 P_{f3}=0.00000

It can be seen from Table I that the system failure probability is not evenly distributed to each failure mode, and it is greater than all the modal failure probabilities. Therefore, design by imposing probability limits only on individual modes may not yield desirable system reliability. For this specific problem, the vertical deflection limit state is most dominating, and the yielding failure mode has a probability of occurrence of zero, which means that it is not affecting the design. Design considers only g1 and g2 yields the same results as that when all three failure modes are considered.

Summary and Closing Remarks: A new reliability-based design approach that applies for both component and system reliability problems is proposed in this paper. The *g*-functions are used as intermediate response functions in the approximation of the system reliability function. Random variables and design variables are separated through a simple transformation, and each *g*-function is approximated around the current design point and its current most probable points (MPP) with respect both the design variables and the random variables. New design is then obtained using the approximate *g*-functions. This procedure is repeated until convergence of design is achieved or a satisfactory design is obtained. The structural system reliability is estimated using an efficient importance sampling technique that uses approximate quadratic g-functions with adaptive curvatures and positions. The efficiency and effectiveness of this method is demonstrated by designing a cantilever beam under lateral and vertical random loading against three failure modes: excessive lateral and vertical deflection and yielding. The same beam is also designed using the conventional approach, which requires 5 times more structural analysis than the proposed method. The results also show that different failure mode has different probability of occurrence at the optimal design. It is therefore important to consider a problem with multiple failure modes as a system problem and not as a problem with many independent and equally import failure modes.

Acknowledgement: The support of the Southwest Research Institute Internal Research and Development Project 06-9970 is gratefully acknowledged.

References:

1. Numerical Optimization Techniques for Engineering Design: With Applications, Vanderplaats, G.N., McGraw Hill, NY, 1984.

2. Structural Reliability - Analysis and Prediction, Melchers, R.E., Ellis Horwood Ltd/ John Wiley, New York, NY, 1987.

3. "Reliability Based Optimization: A Safety Index Approach," Nikolaidis, E., and Burdisso, R., *Computers & Structures*, Vol. 28, No. 6., pp. 781-788, 1998.

4. "Integrated Analysis and Design in Stochastic Optimization," Maglaras, G.K., and Nikolaidis, E., *Structural Optimization*, Vol. 2, pp. 163-172, 1990.

5. "A g-function Based Reliability-Based Design Methodology Applied to a Cam Roll System," Wang, W., Sui, P.C., and Wu, Y.-T., Proceedings of the 2nd *World Conference in Integrated Design & Process Technology*, Austin, Texas, December 1-4, 1996, L.L. Esat et.al., eds., Vol. 3, pp. 70-73, 1996.

6. "Efficient Probabilistic Design Optimization Method by Converting the Reliability Constraints to Equivalent Deterministic Constraints," Wu, Y.-T., and Wang, W., Presented at the 2nd *World Conference in Integrated Design & Process Technology*, Austin, Texas, December 1-4, 1996, L.L. Esat et al., eds., Vol. 3, pp. 55-63, 1996.

7. "Reliability-based Structural Optimization: A Simplified Safety Index Approach," Reddy, M.V., Grandhi, R.V., and Hopkins, D.A., *Computers and Structures*, Vol. 53, No. 6, pp.1407-1418, 1994.

8. "Computational Methods for Efficient Structural Reliability and Reliability Sensitivity Analysis," Wu, Y.-T., *AIAA Journal*, Vol. 32, No. 8, pp. 1717-1723, 1994.

9. DOT (Design Optimization Tools), VMA Engineering, Version 2.05, 1990.

CORRELATION PROCEDURES FOR FATIGUE LIFE DETERMINATION.

Raed Kouta, *Assistant Professor*
Université Technologique de Belfort-Montbéliard
Laboratoire d'Ingénierie de Produit et d'Ergonomie
Rue du château - 90010 - SEVENANS
☎ : 33 3 84 58 31 74 Fax : 33 3 84 58 3030

Daniel Play, *Professor*
Institut National des Sciences Appliquées de Lyon
Laboratoire Conception et Analyse de Systèmes Mécaniques
20, Av. Albert EINSTEIN, Bât. 113
69621 - VILLEURBANNE Cedex FRANCE
☎ : 33 4 72 43 82 48 Fax : 33 4 78 94 60 98

Abstract: Fatigue life determination asks for a fine definition of load spectra applied to parts of vehicles. Significant load spectra must be defined from experimental data. Global shapes of loading distribution curves are obtained from Gram-Charlier-Edgeworth mathematical developments. Further considering the physical aspect of material fatigue behaviour, a local definition of extreme loading is introduced. Then, frequency analyses are linked to the distribution of amplitudes. 200 practical recordings were analysed. Equivalent load spectra are compared to S-N curves and correlation between real loading of vehicle parts in service and laboratory test loading are obtained.

Key Words: Signal treatment, Fatigue Analysis, Correlation Procedures,

INTRODUCTION: To day, fatigue tests on vehicles always take a big place in prototype qualifications. Those tests are necessary to adapt vehicle, performances to customers requirements. Market studies and behaviour analysis of users define how vehicles are used in various environments. Tests were performed on vehicles which were driven on different test tracks [11, 12]. Each track is subdivided proving ground sections whose critical aspects constitute severity indexes (gravely ground, pothole section, concrete bumps, Belgian blocks section, etc...) [3]. Tests procedures done on a reduced distance should be able to reproduce real use of vehicles. It is wellknown that several runs on each section are needed to ensure a good reproducibility of tests. However, number of runs has to be limited because of cost limitations and test data storage conditions. Consequently, a minimum of runs and accurate length of tracks must be defined.
Load spectra derived from classical statistics analysis of test recordings are often multimodal. Hence, each track must be analysed section by section. Indeed, each section gives its own random and frequency characteristics. This study deals with the analysis of random loading applied to mechanical parts of vehicles. Statistical approach and morphology of signals will be considered. Usually, signal treatments are based on numerical counting (overcrossing level, rain flow counting, etc...) and each counting technic presents a particular treatment. In this paper, signals are considered from a statistical points of view. Global shapes of signal distributions will be defined as well as

the local characteristics of the signals. The study goes on a better identification of measured signals with a non-gaussian approach. Further, severity modes of test track will be analysed and risk of overtaking stops for mechanical parts will be linked to the running distance. Correlation's based on mechanical loading between test tracks and test sections give a better classification of test roads. This is done in relation to severity characteristics of the test tracks transmitted to parts of vehicles. Severity indexes will be defined. Moreover, a method will be proposed to make extrapolation of fatigue life of mechanical parts from section-track test mileage to a real customer use that corresponds to a long length in kilometres.

MODELIZATION OF STATISTICS DISTRIBUTIONS: As the approach is based on probability of signal amplitude overcrossings, elementary statistics becomes no longer convenient. A more accurate mathematical approach must fit the different curves of load spectra. Spectrum momentum's of second, third and fourth order are thus introduced. The skewness factor β_1 represents asymmetry while the kurtosis β_2 corresponds to the flattening of the curve. The momentum's μ_k are calculated from a set of x_i values obtained at times t_i and the average value $\bar{x}$. The table I shows eight values obtained on four mechanical parts for two test sections. Note that the measurement were made with mechanical or physical limits (rubber stops at the suspension displacements,...). The two parameters a, b give the limits. Remember that a gaussian distribution gives β_1=0.0 and β_2=3, the results are far from these classical values.

Amplitude variations of signals measured on mechanical parts correspond to the dynamic response of the vehicle submitted to excitations given during tests along tracks. Small fluctuations cannot be distinguished from global distributions while they also contribute to the fatigue behaviour. Thus, it becomes necessary to study more precisely the global and the local distribution of the random signal amplitudes.

Load	A (mm)		B (mm)		C (mdaN)		D (ms^{-2})	
a	-100		-100		-400		-50	
b	100		100		400		50	
Section	GT	BR	GT	BR	GT	BR	GT	BR
min	-93	-79	-42	-72	-410	-274	-98	-72
max	97	63	56	71	503	367	117	95
$\bar{x}$	-0.4	1.15	-1.4	0.43	2	2.7	0.0	0.00
σ	21	16.0	8	18.0	91	78.0	9.8	10.
β_1	0.367	0.087	-0.08	0.244	0.03	0.757	0.22	0.063
β_2	4	4	5	4.12	4.4	4.11	12	6.00

Table I: Values of statistics parameters.

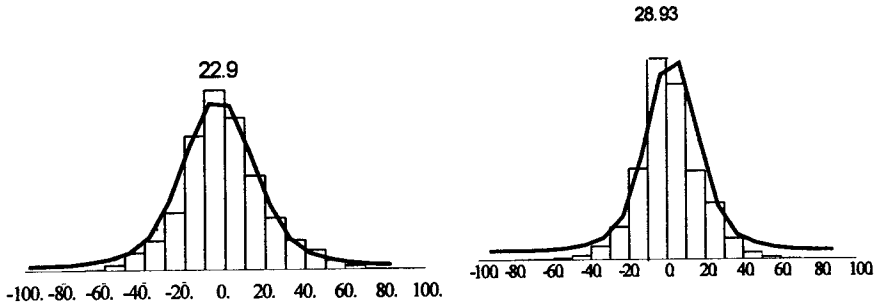
A : Displacement of right front suspension,
B : Displacement of left front suspension,
C : Torque on front antiroll bar,
D : Vertical acceleration of rear axle,
GT : Gravely track,
BR : Bad road conditions.

a = lower limit,
b = higher limit;
$\bar{x}$: Average value,
σ : Standard deviation;
min : Minimum; max : Maximum;
β_1 skewness; β_2 kurtosis

Model of global statistics distribution: The serial development of Gram-Charlier and Edgeworth [4] allows to define a probability density function g(x) as follow :

$$g(x) = \exp\left[\sum_{j=1}^{j\leq 4} \kappa_j \left\{(-D)^j / j!\right\} \alpha(x)\right], \quad \left| \begin{array}{l} D=d/dx,\ \alpha(x) : \text{Gauss law,} \\ kj : \text{cumulative terms } (k_3=(\beta_1)^{1/2},\ k_4=\beta_2-3) \end{array} \right. \quad (1)$$

The curve fitting is obtained through the adequation test χ^2. In this way, frequency diagrams are replaced by mathematical laws (Fig. 1). The visualisation corresponds to the displacements of the right front suspension for two conditions of use (GT: gravely track, BR : bad road conditions).



GT : Gravely Track **BR : Bad Road conditions**

Fig. 1: Frequency diagrams and density probability functions
(A : displacements of the right front suspension)

Model for local statistics distributions: Four cases of extreme values can be observed during amplitude variations of signals (Fig. 2):
- maximum positive value with a preceding positive gradient (peak > 0)
- minimum negative value with a preceding negative gradient (valley < 0)
- maximum negative value with a preceding positive gradient (peak < 0)
- minimum positive value with a preceding negative gradient (valley > 0)

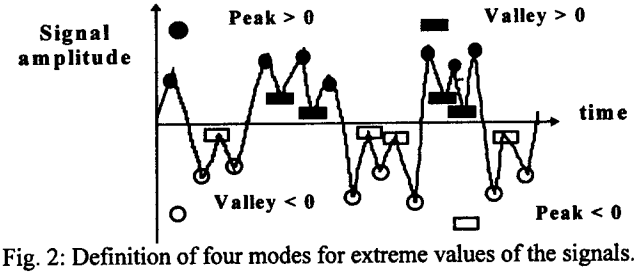


Fig. 2: Definition of four modes for extreme values of the signals.

Four statistic samples can be obtained as well as four Weibull distributions whose shape factors are estimated from the studied samples [5]. Weibull parameters are connected to the studied phenomena :
- gap (d) defines the threshold of amplitude when the phenomenon appears,
- shape factor (β) shows the density of amplitude occurrences,
- scale parameter (η) indicates the extended area of the distribution.

The probability density function f(x) and the cumulative distribution function F(x) are defined : $f_X(x) = \frac{\beta}{\eta}(\frac{x-d}{\eta})^{\beta-1} \exp(-(\frac{x-d}{\eta})^{\beta})$; $F_X(x) = 1 - \exp(-(\frac{x-d}{\eta})^{\beta})$ (2)

In the same time, an amplitude rate can be defined specially when the sample at large amplitudes is small (Fig. 3). $\lambda(x)=f_X(x)/(1-F_X(x)$, (3)

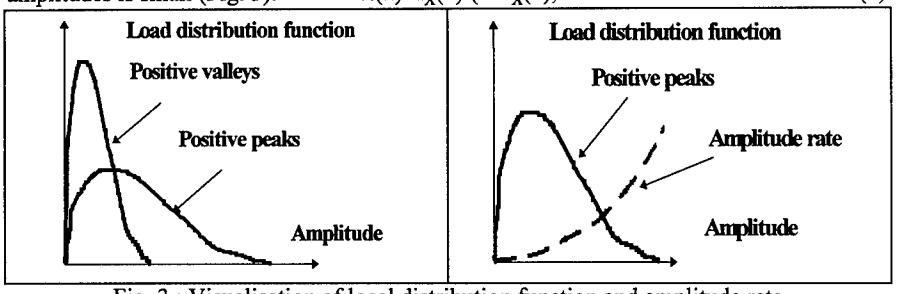


Fig. 3 : Visualisation of local distribution function and amplitude rate

Risk Coefficient and Severity Coefficient: The measurements give obviously a sample among all the loading. Consequently, the observed extreme values are only the realisation of random variables. Thus a zoom of the ends of the global distribution (Fig. 4) allows to introduce theoretical models (4) for laws corresponding to the minimum and maximum [10]:

law for minimum: $F_{min}(x)=1-[1-F_X(x)]^n$ | law for maximum $F_{max}(x)=[F_X(x)]^n$ (4)

$f_{min}(x)=DF_{min}(x)$ | $f_{max}(x)=DF_{max}(x)$

F : cumulative distribution function; $F_U(u)=Proba(U<u)$;

n: sample; f: probability density function; D=d/dx

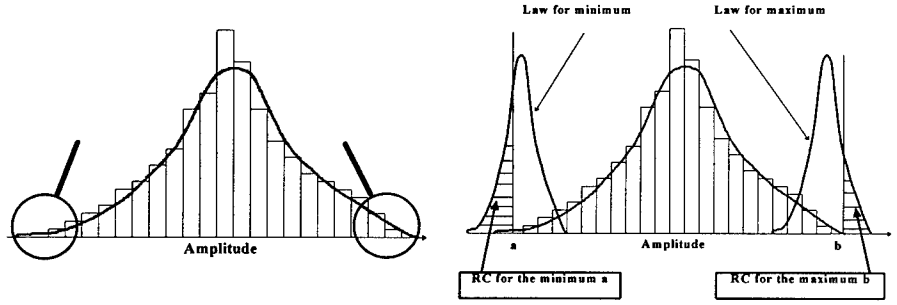


Fig. 4: Zoom at the ends of the amplitude distribution and risk coefficient visualisation

A Risk Coefficient (RC) of overcrossing of a given damaging amplitude (a or b) is then obtained with the area under the curve. The same analysis can be performed with the local distribution and a Severity Coefficient (SC) is then introduced (Fig. 5). The table II sums up the numerical values for the four loading and the two sections. Note that the risk coefficients are practically equal to zero while the severity coefficients show significant differences. Those results point out the possibility to evaluate the performance of one section and to make comparisons among sections [6].

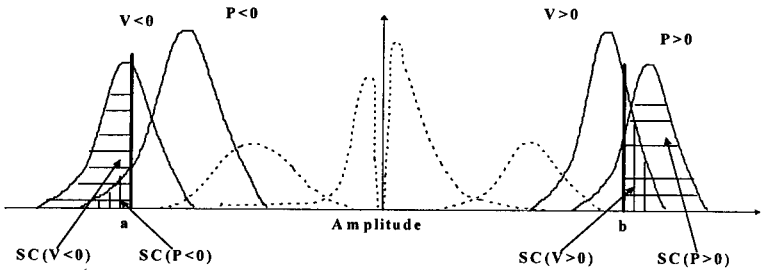


Figure 5: Density probability functions at the ends of local distributions and severity coefficient visualisation

RC	Minimum		Maximum		SC	V < 0	P < 0	V > 0	P > 0	V < 0	P < 0	V > 0	P > 0
Section	GT	BR	GT	BR	Section	GT				BR			
A	0.054	0.0	0.075	0.0	A	0.135	0.025	0.398	0.714	0.054	0.02	0.0	0.0
B	0.0	0.0	0.0	0.0	B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	C	0.271	0.023	0.04	0.21	0.0	0.04	0.064	0.15
D	0.0	0.0	0.0	0..0	D	0.633	0.371	0.233	0.681	0.0	0.233	0.0	0.02

Table II : Risk and Severity Coefficients

IMPROVED AMPLITUDE SPECTRA FOR FATIGUE DETERMINATION:

Frequency approach for counting the amplitudes : The probability to overcross an amplitude level depends on the actual gradient of the signal. In the same way the probability to found an extreme value depends on the actual radius of curvature of the signal. The coupling between amplitude event and signal shape is made with the help of first and second derivative of the signal.

The probability to overcross an intermediate level of amplitude is first considered. This is done for different classes of amplitudes (horizontal bands on the figure 6). Two different signals with the same number of events per class give two levels of severity. An improved counting is proposed in order to obtain the frequency of occurrence for each amplitude class and taking into account the average value of the slope of the signal for the considered class. It is based on a theoretical development of overcrossing level [7, 9, 1, 2]

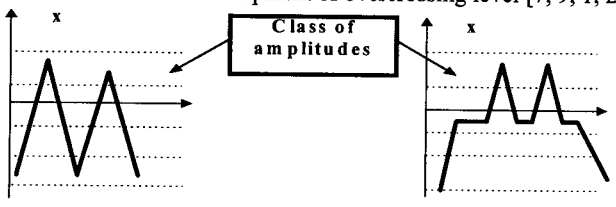


Fig. 6 : Two different signals for the same amplitude counting

$$N_{\alpha}(c_k) = g_X(\alpha) \cdot E\left\{|x'| / x \in \left[\alpha - \frac{\Delta c}{2} ; \alpha + \frac{\Delta c}{2}\right]\right\} \quad (5)$$

$$N_0(c_k^*) = g_X(0) \cdot E\left\{|x'| / x \in \left[-\frac{\Delta c}{2} ; \frac{\Delta c}{2}\right]\right\} \quad (6)$$

Where
Nx: improved counting of amplitude overcrossing for one class of amplitude with an average amplitude (x),
$g_X(\alpha)$: the Gram-Charlier-Edgeworth probability density function,
E{.} : mean value function, ck : k^{eme} class; Δc : width of one class,

The formula (5) defines a number of amplitude overcrossing for one class of amplitude. The formula results in the weighting of the likely number obtained from statistics analysis by the average value of the signal gradient.
In the same way, the probability to reach an extreme value $f_M(\alpha)\, d\alpha$ can be defined in the amplitude window $\alpha<x(t)<\alpha+d\alpha$. This distribution depends on the distribution $g_{xx'x''}(u,v,w)$ that links the signal (x) and their two first derivatives. Theoretical developments can be made [9] but a first approximation $g_{xx'x''} = g_{x'} * g_{xx''}$ corresponds to a nil correlation between the signal and the first derivative. Thus, a bivariable Pearson-Sagrista distribution [8] is written :

$$g_{XX''}(u,w) = k(h^2 + au^2 + 2buw + cw^2)^n \mathrm{Exp}\left(-\frac{1}{2}(au^2 + 2buw + cw^2)\right) \qquad (7)$$

$$\text{and } k = \sqrt{ac - b^2} \Big/ \pi e^{h^2/2} \int_{h^2}^{\infty} u^n e^{-u/2} du$$

This function gives with its momentum when n=2 the function of the signal and of the second derivative., $f_M(\alpha)$ becomes :

$$f_M(\alpha) = \frac{I_{ng}\, e^{-x^2/2}}{V(x')\sqrt{2\pi}} \left[\left\{ \left(\frac{1 + I_{ng}^2}{1 - I_{ng}^2}\right)x^4 + \left(\frac{4 - I_{ng}^2}{1 - I_{ng}^2}\right)x^2 + \frac{3I_{ng}}{\sqrt{1 - I_{ng}^2}}x + 8 \right\} e^{-\frac{I_{ng}^2 x^2}{2(1 - I_{ng}^2)}} + \frac{I_{ng}}{\sqrt{1 - I_{ng}^2}} x(x^4 + 4)\sqrt{\frac{\pi}{2}}\,\mathrm{erf}\left(\frac{I_{ng}\, x}{\sqrt{2(1 - I_{ng}^2)}}\right) \right] \qquad (8)$$

With a=V(x), b=V(x'), C=V(x''), $x=(\alpha - \bar{x})/(V(x))^{1/2}$, and h=0

The figure 7 shows the influence of the signal irregularity factor I_{ng} on the distribution of extreme values.

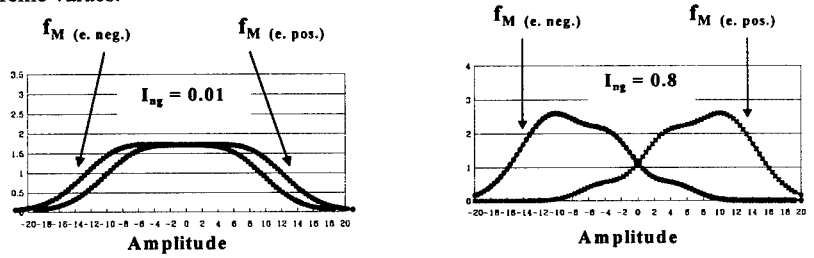


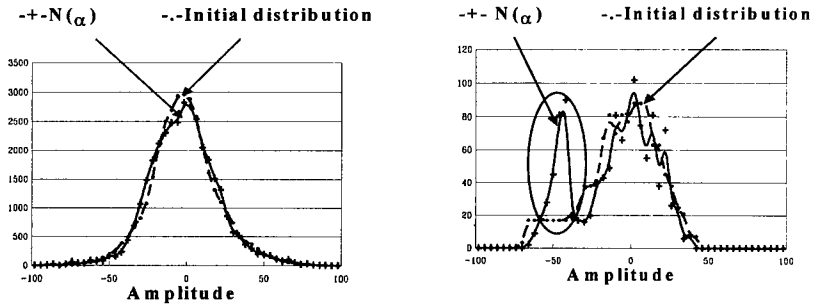
Fig. 7 : Non gaussian distribution of extreme values ($f_M(\alpha)$)
(Sagrista model n=2, e.= extreme, neg.=negative, pos.=positive)
in relation with irregularity factor

The set of curves corresponds respectively to positive and negative extreme values. A low irregularity factor corresponds to a random signal with both a lot of intermediate fluctuations between two consecutive extreme values and a lot of intermediate average values. In the opposite case, a large irregularity factor corresponds to one average value and a lot of variations of extreme values.

Here also, the weighting of the probability density function of the extreme values, is obtained with the average value of signal curvature for each class of amplitudes :

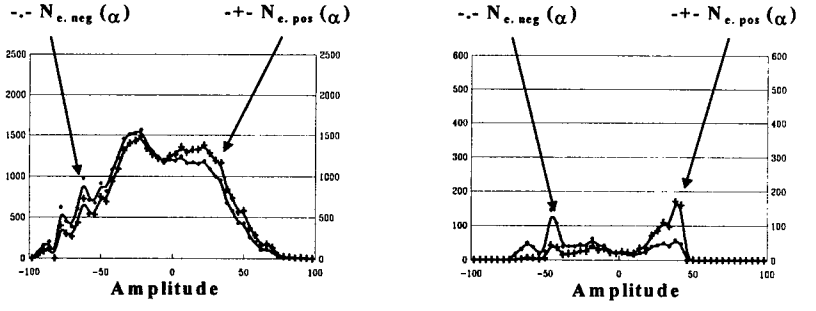
$$N_e(c_k) = f_M(\alpha) \cdot E\left\{|x''| / x' = 0 \; x \in \left[\alpha - \frac{\Delta c}{2} ; \alpha + \frac{\Delta c}{2}\right]\right\} \quad (9)$$

Where : N_e : improved counting of extreme values for one class of amplitude,
$f_M(\alpha)$: the extreme value probability density function, $E\{.\}$: mean value function,
ck : k^{eme} class; Δc : width of one class,



a) GT b) CB

Fig. 8 : Improved counting results for amplitude overcrossing
(Gravely Track, Concrete Bump)



a) GT (I_{ng} =0.08) b) CB (I_{ng} = 0.6)

Fig. 9: Improved counting results for extreme values of the signal
(Gravely Track, Concrete Bumps)

Variations of average slopes and curvatures for signals obtained on GT and BR tracks are small as only one frequency of mechanical excitations occurs. Consequently, small changes appear on the improved counting curves (Fig. 8a and Fig. 9a). However, other tracks can give large differences. For example, a concrete bumps track produces large

values of average signal slopes for the intermediate class of amplitudes. Particularity for the negative amplitudes, the initial distribution (Fig. 8b.) presents a plateau (circle zone on the figure). A part of the signal variations is close to sinusoidal variations. Slopes become larger for the intermediate amplitude values and consequently the real number of amplitude overcrossing must be increased.
Further, signal curvatures are seen larger at higher amplitudes and improved counting is more effective in this case. Finally a small I_{ng} gives two closeness curves while a large I_{ng} produces different curves.

Definition of improved spectra: Measurements on mechanical parts give specific information depending on both on the dynamic mechanical behaviour and road excitations. Consequently, a fine observation of each part of signals allow to introduce specific separate effects. Now from a statistical point of view, combination of these effects with give improved spectra.
Remember that a classical spectra of loads is obtained with specific counting methods [13, 14]. Results are largely dependent on the specific assumptions of the counting methods. A random variable X_{class} and achievements of this variable $x_i(t_i)$ are considered. For example in the case of amplitude overcrossing, the frequency diagram represents the numbers of amplitude overcrossing predefined amplitude level. Thus, a classical probability function is obtained Proba(X_{class}<x). In this study, signal variations are followed step by step during time rather than to get a sightseeing view of the overall signal. Consequently, a new probability function is obtained through a combination of classical and local probability functions: Proba(X_{class}<x); Proba(X_{max}<x), Proba(X_{min}<x); Proba($X_{/slope>0}$<x), Proba($X_{/slope<0}$<x), Proba($X_{/curvature<0}$<x) et Proba($X_{/curvature>0}$). If independance of the previous random variables is assumed, improved spectra can be introduced. The first global improved spectrum (GIS) concerns 6 specific spectra :
- Gram-Charlier-Edgeworth probability density function ($n*g_X(x)$),
- the law of minimum ($n*f_{min}(x)$ and the law of maximum ($n*f_{max}(x)$),
- improved counting of amplitude overcrossing ($N\alpha$),
- two improved counting of extreme values (Ne).

$$GIS_{Le}(\alpha) = \frac{1}{6}\left[(n.g_X(\alpha)) + (n.f_{min}(\alpha)) + (n.f_{max}(\alpha)) + N_\alpha + N_{e.neg.}(\alpha) + N_{e.pos.}(\alpha)\right] \quad (10)$$

n = scale of the sample; Le = Length of a tested track (km); neg. = negative; pos. = positive

The second spectrum is concerned with eight local aspects of signal and is called local improved spectrum (LIS):

$$LIS_{Le}(\alpha) = \frac{\lambda(\alpha)}{8}\left[\begin{array}{l} n_{v<0}.f_{v<0}(\alpha) + n_{p<0}.f_{p<0}(\alpha) + n_{v>0}.f_{v>0}(\alpha) + n_{p>0} \times f_{p>0}(\alpha) + \\ n_{v<0}.f_{max.v<0}(\alpha) + n_{p<0}.f_{max.p<0}(\alpha) + n_{v>0}.f_{max.v>0}(\alpha) + n_{p>0}.f_{max.p>0}(\alpha) \end{array}\right] \quad (11)$$

$n_{v<0}$= scale of the sample for the negative valleys;
$n_{v>0}$= scale of the sample for the positive valleys;
$n_{p<0}$= scale of the sample for the negative peaks;
$n_{p>0}$= scale of the sample for the positive peaks.

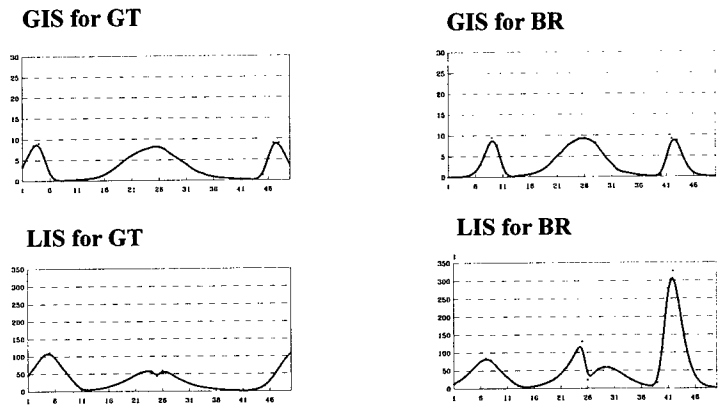


Fig. 10 : Global and local improved spectra (A displacement of right front suspension).

The figure 10 shows that GIS and LIS distributions can get different shapes and the probability of overcrossing maximum amplitudes is larger for the Gravely Track case.
The previous spectra are defined from test tracks and obviously are used to define fatigue life of vehicles working during a long length in kilometres. Consequently, the previous spectra must be weighting with a proportionality factor between the large number of kilometres K for customers, and the small numbers considered during tests Le :

$$GIS_K(\alpha) = \frac{K}{Le} GIS_{Le}(\alpha) \qquad LIS_K(\alpha) = \frac{K}{Le} LIS_{Le}(\alpha) \qquad (12)$$

SYNTHESIS AND CONCLUSION: Improved spectra can be compared to typical S-N* curves and classical damage calculation can be achieved. In this study, a theoretical S-N* model was considered $N^* \approx |\alpha|^p$ where p=5. The classical Palmgren-Miner cumulative damage summation was also considered

- for the GIS $d(SAG_{Le} ; \alpha) = SAG_{Le}(\alpha)/ N^*(\alpha)$
- for the LIS $d(SAL_{Le} ; \alpha) = SAL_{Le}(\alpha)/ N^*(\alpha)$

where $N^*(\alpha)$ corresponds to the number of cycles given by the S-N* curve at the α amplitude.
Consequently, the summation of the elementary damage d gives a pseudo-damage D. The results are gathered in the table III for two tracks.

Damage		*Gravely track (GT)*	*Bad Road (BR)*	Ratio = GT/BR
GIS	A	0,1239	0,0306	4
	B	0,001145	0,0004	3
	C	0,0622	0,02586	2
	D	0,00064	0,0006	1
LIS	A	17,2e-4	1,12E-03	2
	B	7,76E-05	8,15E-05	1
	C	3,82E-04	2,23E-04	2
	D	1,83E-04	1,73E-04	1

Tableau III: Pseudo-damage values calculated from global and local improved spectra

Note that the damages calculated with rain-flow counting and overcrossing level methods gave ratios practically equal to 1. Note also that experimental evidence from test results rates the GT section, more damaging than the BR section. In the table III the ratio GT/BR for the four mechanical parts is often larger than one. From a global point of view in a first approach, the numerical results show that a mean coefficient of 2 can be deduced: a test performed on gravely track is twice as sever as test on bad roads. Obviously, better definitions of test severity will be introduced for particular mechanical parts.

In conclusion, fatigue life determination is based on the knowledge of both material constitutive laws and loading spectra. The material studies have received large parts of efforts but comparatively the efforts allocated to actual loading determinations stay small. Numerical approaches of dynamic responses of mechanical systems can be achieved now. However, it requires a lot of computing times and the entries of calculations are not wellknown. Consequently, experimental approaches are always in practice. The study presented in this paper dealed with the understanding of loading signals applied on mechanical parts in respect to intrinsic material fatigue life behavior. Shape indexes of the global loading distributions become no longer sufficient to take into account all the intermediate loading. Moreover as the number of loading controls the fatigue damage cumulating, the frequency character of loading signals was introduced through slopes and curvatures of signals. In this way, global and local improved spectra are introduced to give a better description of the mechanical part loading.

References and Bibliography:

[1] : R. KOUTA, "Méthodes de corrélation par les sollicitations des pistes d'essais de véhicules". PhD Lyon, France, N° 94ISAL 0096, nov. 1994.

[2] : R. KOUTA, J. FAUCHON, D. PLAY, "Procédures de corrélation entre essais accélérés et conditions réelles d'exploitation, concernant un véhicule industriel", 9th international conference on Reliability and Maintainability, European Saftey and Reliability conference, λμ 9, La Baule, France, June 1994.

[3] : W. A. GROSS Jr ; "Aberdeen proving ground facilities for testing military vehicles", Soc. of Auto. Eng. Inc. Mid-year Meeting, Detroit, Mich., May 1970, 25 p., SAE 700523.

[4] : M. KENDALL, A. STUART : "The advanced theory of statistics", Griffin, G.B. 1973

[5] : J. BANKS, J.S. CARSON : "Discret-event system simulation", Prentice-Hall, INC., Englewood Cliffs, New Jersey, 1984.

[6] : N.C. JACKSON ; "Pavement performance - A state DOT perspective", Truck and Bus Meeting and Expo. Indianapolis, Indiana, Nov. 7-10 1988, pp. 5-15.

[7] : A. PREUMONT ; "Vibrations aléatoires et anayse spectrale", Lausanne: Presses polytechniques et universitaires romandes, 1990, 343 p.

[8] : S. N. SAGRISTA ; "On a generalization of PEARSON's curves to the two-dimensional case", Trabajos de Estadistica, 1952, No. 3, pp. 273-314.

[9] : J. R. RICE, F. P. BEER, P. C. PARIS ; "On the Prediction of Some Random Loading Characteristics Relevant to Fatigue", Acoustical Fatigue in Aerospace Structures. Proc. of the Sec. Int. Conf. Dayton, Ohio, April 29-May 1, 1964, pp. 121-144.

[10] M.R.LEADBETTER, G. LINDGEN, H. ROOTZEN : Extremes and related properties of random sequences and processes", Springer series in statistics, 1980

[11] : V. GRUBISIC, Détermination of load spectra for design and testing, Int. J. Of vehicle design, Vol 15, 8-26 p, 1994.

[12] : E. LESSE, L MULLIN, Role of fatigue analysis in the vehicle test simulation laboratory, SAE Technical paper series, n° 91016, 57-68p, 1991.

[13] : H. D. BERNS ; "Field Service History Analysis For Ground Vehicles", Peoria, Illinois, .Soc. of Auto. Eng. Inc., Earthmoving Industry Conf., 1975, 10 p., SAE 750553.

[14] : G. M. VAN DIJK ; "Statistical load data processing", 6th Int. Conf. Aircraft Fatigue Symp., Miami, FLORIDE, 1971, NASA SP-309, pp. 565-598.

COMPARATIVE STATISTICAL ANALYSIS OF FATIGUE CRACK GROWTH DATA

J.D. Baldwin*, C.A. Paul**

*School of Aerospace & Mechanical Engineering
University of Oklahoma
Norman, Oklahoma 73019

**Flight Dynamics Directorate
Wright Laboratory
Wright-Patterson AFB, Ohio 45433

Abstract: The differences in fatigue crack propagation behavior in 7075-T6 aluminum alloy between specimens damaged by corrosion and those in a baseline (noncorroded) condition are examined using a statistical hypothesis test. Middle tension plate specimens were corroded on one side in the laboratory and subjected to constant load amplitude (K-increasing) tests in dry and moist air environments. The results show that there is a statistically-significant increase in fatigue crack growth rates in the corroded material for stress intensity values based on nominal thickness. When the specimen thicknesses were reduced to account for corrosion material loss, the difference disappeared in most of the instances examined here. In one case, the thickness correction was found to predict fatigue crack growth rates that were *slower* in the corroded material. This anomalous result suggests that full thickness corrections may not be appropriate at higher stress intensity ranges.

Key Words: Corrosion fatigue; Damage tolerance; Fatigue crack propagation; Prior corrosion; Structural integrity

Introduction: In fracture mechanics-based structural durability analysis, one of the important material response parameters is the fatigue crack growth rate (FCGR), characterized by da/dN versus ΔK data. Here a is the crack length, N is the number of load cycles and ΔK is the stress intensity factor range for the load cycle. Although the fatigue crack growth rate data for a material can be presented in the form of an empirical equation (e.g., the Paris equation), it is typically given graphically. As with other types of fatigue data, crack growth experiments display considerable scatter. Usually the variability in FCGR data is viewed as a nuisance and little account is taken of it. Attempts have been made, however, at incorporating the variability into probabilistic crack growth models. Specifically, Virkler, et al. [20] and Ghonem and Dore [8] developed experimental data that clearly illustrates the crack growth rate variability and scatter. Recent analyses by Yang and Manning [21] and Maymon [12] illustrate current developments in stochastic crack growth analysis. The data scatter also becomes of great interest, however, when one is faced with the task of comparing two sets of FCGR data to detect whether they are substantially equal or not. Such an analysis is particularly important when differences in cracking behavior are anticipated due to experimental treatments such as surface condition, environment and stress ratio.

The analysis developed here has come about within the context of the U.S. Air Force Aging Aircraft program. The USAF's fleet of cargo and tanker aircraft are being operated well beyond their original design life and structural damage mechanisms that were not considered in the design process, specifically corrosion, are becoming of increasing interest in assessing the continued durability of these aircraft. It is now well-known that the average ages of commercial and military aircraft fleets are increasing, prompting an intense examination of the tradeoffs between economic efficiency and airworthiness. As the fleets age, corrosion damage and its impact on airframe durability become of particular concern. As Schutz noted [19], fatigue strength and corrosion parameters combine in a

synergistic manner that has not yet been completely described. The technical challenges facing operators of aging aircraft came into sharp focus in April 1988 when Aloha Airlines Flight 243 experienced a catastrophic failure of the forward fuselage. Although the cause of the Aloha structural failure was formally given as multiple site damage, there was a component of corrosion damage involved. Corrosion-induced aircraft accidents have occurred throughout the history of aviation. Campbell and Lahey [1] reported that, since 1927, there had been over 60 accidents worldwide at least partly attributable to corrosion damage. In a more recent study, Hoeppner, et al. [10] update Campbell's statistics, reporting that since 1975 there have been nearly 700 domestic incidents and accidents in which corrosion was at least a contributing factor.

Hendricks [9] estimated the cost of a major commercial aircraft overhaul at $2-20 million. Compared with the approximately $50-100 million required for a new aircraft, the economic incentive for operating older planes is apparent. For operators of large fleets, however, the economic burden of keeping aging aircraft operational is becoming prohibitive. In 1994, a U.S. Air Force study showed that the maintenance costs of repairing corrosion damage alone had reached the $1-3 billion per year level [2]. Since the Aloha accident, the issues involved in operating aging aircraft fleets have received increased attention from agencies such as the U.S. Air Force, U.S. Navy, National Aeronautics and Space Administration and Federal Aviation Administration. It is important to realize that corrosion damage, one of the primary sources of aircraft structural degradation, is not accounted for in the design structural analysis, and thus in the inspection intervals, of any major aircraft system.

The U.S. Air Force's fleet of nearly 700 C/KC-135 tanker aircraft, first designed in the mid-1950's, now has an average age of more than 30 years with none newer than 24 years old [2]. By virtue of its vintage, the C/KC-135 was designed before the damage tolerant design philosophy was mandated for all USAF aircraft and therefore, does not include features such as crack arrestors commonly found in more recent designs. Corrosion of the aluminum fuselage panels and wing skins has been observed with increasing regularity, particularly in lap joints and around fastener holes. In fact, corrosion has proven to be an insidious enemy, often found only after disassembly of the structure, despite the use of nondestructive inspection systems. These tendencies are observed not only in the C/KC-135 fleet, but in other aging weapons systems as well. Because of the prohibitive cost to replace the fleet, these aircraft are considered a national asset and, as such, will be expected to serve well into the next century [11]. In fact, it is anticipated that the C/KC-135's will be one of the last of the current USAF transport platforms to be retired. Because there is no plan to replace these aircraft in the near future, the effect of corrosion on structural integrity must be quantified and incorporated into the fleet maintenance procedures.

Regarding corrosion, the relevant characteristic of the operation of such aircraft is that they spend most of their time outdoors on the ground and only a few hundred hours per year in flight. Therefore, unlike the classical case of corrosion fatigue, where corrosion and fatigue processes evolve concurrently, to a first approximation the corrosion and fatigue damage processes can be assumed to be uncoupled in the case of military transport aircraft. Only recently have fracture mechanics-based experiments designed to quantify the fatigue response of corroded metal begun to appear. Chubb, et al. [3,4,5] examined the effect of exfoliation corrosion on the fatigue crack growth rates of 2024-T351 and 7178-T6 alloys. Chubb's data indicates that at low ΔK levels, the corroded material experienced crack accelerations of up to five times over noncorroded material. Scheuring and Grandt [17,18] used 2024-T3, 7075-T6 and 7178-T6 materials taken from retired C/KC-135 aircraft corroded in service. Scheuring notes that the lightly corroded material shows little difference in crack growth behavior compared to noncorroded material. Their data also shows that the acceleration in crack growth tends to diminish as the stress ratio R increases. While these two research teams report valuable, although limited, results, the conclusions drawn are based on

qualitative comparison of the data. Lacking a statistical analysis, it is difficult to separate the effects of the test parameters such as corrosion damage level, stress ratio and test environment.

Because of the lack of a suitable body of fatigue data on aluminum alloys damaged by prior corrosion, a material characterization program was begun by the USAF to develop some basic fatigue crack growth rate data on specimens with laboratory-grown corrosion damage. The goal of the program was to compare the crack growth behavior of corrosion-damaged material with the baseline, noncorroded material. From that data, conclusions would be drawn about the impact of corrosion damage on fatigue performance; subsequent policy decisions regarding continued airworthiness and any necessary modifications to inspection intervals and procedures would then be based on a suitably large body of experimental data.

The statistical analysis reported here was conceived to examine two issues relating to the fatigue cracking behavior of metal damaged by prior corrosion. First, the existence of any corrosion-based crack acceleration had not previously been addressed rigorously. As noted above, Chubb, et al. [3,4,5] reported crack accelerations of up to five times in corroded material. On the other hand, Scheuring and Grandt [18] observed little or no differences in the *da/dN* versus ΔK data for "lightly" corroded material. It is important to note that both of these conclusions were drawn without resorting to statistical analysis and, rather, were based on visual observation of the data. Given the potential economic and safety ramifications of an incorrect assessment of the relative performance of baseline and corroded material, such subjective analysis is inadequate. The second issue of interest is whether or not modeling the corrosion damage as simply a thickness loss from the specimen is sufficient to describe any differences in crack growth rates; this hypothesis was initially proposed by Doerfler, et al. [6]. If a significant difference persists after the thickness correction, one might be alerted to the presence of a secondary process effecting crack growth rates, perhaps acting on a microstructural scale. Again relying on a subjective assessment, Scheuring and Grandt reported that correcting their ΔK values for measured thickness loss did not alter the conclusion that there was no effect due to corrosion. In cases such as these, a statistical analysis of the data will allow us to address these questions in a rigorous way, avoiding the subjective nature of visual data inspection.

In this paper we describe a statistical analysis that has been used to address the equality of two or more fatigue crack growth rate experiments. We begin with the body of FCGR *da/dN* vs. ΔK data segregated into two groups: the baseline group (i.e., specimens without the experimental treatment of interest) and the specimens subjected to the experimental treatment. Next, for each experiment, a curve is fit through the FCGR data using a polynomial kernel. Using the curve fits, we can then estimate the value of *da/dN* at any value of ΔK of interest for each experiment and compute the mean value of each of the two groups. A Student t-test is used to test the null hypothesis that the mean *da/dN* for each of the two groups is equal. If we are led to reject the null hypothesis, we then conclude that the experimental treatment has an effect on the fatigue cracking behavior of that material. Examples are given illustrating the analysis for aluminum alloy specimens damaged by prior corrosion compared to specimens without corrosion.

Experimental Data: In order to motivate and illustrate the comparative statistical analysis, we will use a series of test data developed in the USAF Round-Robin Corrosion Fatigue Program. The round robin test program was implemented to explore the effect of prior corrosion on fatigue behavior of representative aircraft aluminum alloys by comparing fatigue crack growth rates in corroded and the baseline (noncorroded) specimens. *K*-increasing fatigue crack growth rate tests were conducted according to ASTM E 647 on samples of 7075-T6 aluminum harvested from retired USAF KC-135 aircraft. All specimens were 1.75 inches wide and 0.062 inches thick (nominally) in the center-cracked plate configuration with EDM crack starter notches. In one half of the specimen population, corrosion was induced on one side of the specimen by exposing them to an ASTM B117 3.5% NaCl

fog solution until the damage was deemed to be "severe"; no other *a priori* assessment of the corrosion damage was made. The other half of the specimen population remained in an as-delivered (i.e., baseline or noncorroded) condition. In addition to the material condition (baseline vs. corroded), the additional experimental parameters stress ratio ($R = 0.05$, $R = 0.50$) and humidity (< 15% R.H., > 85% R.H.) were considered. The test matrix illustrating the multiple replicate full factorial design used in the testing is given in Table 1. In what follows, a test series will be considered to be all of the specimens exposed to a given stress ratio-humidity combination. The round robin testing was conducted in five different laboratories, both within and outside the Air Force. Originally, the intent of the test program was to test six replicates of each test series, but organizational difficulties prevented more than four corroded specimens from being tested in any of the conditions.

In all cases, crack lengths were measured using optical microscopes on micrometer slides. The recorded a versus N data were processed into da/dN vs. ΔK using secant interpolation and the stress intensity solution for the ASTM middle tension specimen

$$\Delta K = \frac{\Delta P}{B}\sqrt{\frac{\pi\alpha}{2W}}\sec\frac{\pi\alpha}{2} \qquad (1)$$

In this expression ΔP is the cycling load range, B is the specimen thickness, W is the specimen width, and $\alpha = a/W$. Any data points that exceeded the ASTM-specified maximum plastic zone size were removed from further analysis as invalid.

Since the appearance of the Paris expression summarizing crack growth rates [14]

$$da/dN = C\,\Delta K^m \qquad (2)$$

it has been traditional to show these data on logarithmic coordinates. Therefore, we will use the data in log da/dN - log ΔK form. For the four experimental test series, the crack growth rate data are shown in Figures 1-4. All of the data here is based on using the actual specimen thicknesses making no deduction for corrosion-induced material loss. In this program, two test conditions contributed to the relatively narrow data domain. First, the scope of the data was restricted at low ΔK levels by not being from decreasing K tests. Also, the relatively narrow test specimens could not develop long enough cracks to result in high ΔK's before reaching the plastic zone size limits of LEFM, thus data at high ΔK levels have been restricted. In Figures 1-4, there is certainly the *appearance* of crack acceleration due to prior corrosion, particularly at lower ΔK levels. It is interesting to note that the baseline and corroded data seem to coincide at the highest ΔK levels reached during testing. Although it has not been examined rigorously, this effect is presumably due to the overriding influence of the crack tip driving force as the crack nears instability. The question remains, however, as to whether the corroded material behavior is really different from the baseline, or are the observed differences simply due to experimental data scatter. In the analysis that follows, we develop and demonstrate a statistical hypothesis test that addresses this concern.

Statistical Analysis: In curve fitting equations of the form $y = mx + b$ (e.g., the Paris equation in logarithmic coordinates), the statistics of the fitting parameters m and b can be computed for each data set [7]. A Student t-test can then be run to identify presumably equal data sets. Such an analysis would, in this case, give only a "global" determination in that the overall equivalence throughout the relevant ΔK range is tested. Our goal was to implement a "local" analysis capable of identifying ΔK ranges where the FCGR's were equal. Thus, the polynomial-based global-local method described here, using a "global" curve fit to make "local" estimates, was found to be suitable for our purposes.

The first step in the statistical analysis is to summarize each experimental data set (the crack length versus cycles data for an individual specimen) by a curve fit of the FCGR data. By establishing a

curve fit for each experiment, we can then interpolate between data points and estimate *da/dN* at convenient values of ΔK, rather than at the actual values which will not coincide in every experiment. There have been many different models used to describe this type of data, ranging from the simple original Paris equation, to nonlinear models such as the hyperbolic sine model [13] given by

$$\log\ da/dN = C_1 \sinh\left[C_2\ (\log\ \Delta K + C_3)\right] + C_4 \tag{3}$$

the Weibull model [15]

$$da/dN = e + (\nu + e)\left[-\ln\left(1 - \frac{\Delta K}{K_b}\right)\right]^{1/k} \tag{4}$$

and a three-parameter reciprocal relationship [16]

$$\frac{1}{da/dN} = \frac{A_1}{(\Delta K)^{n_1}} + A_2\left[\frac{1}{(\Delta K)^{n_2}} - \frac{1}{\left(K_c(1 - R)\right)^{n2}}\right] \tag{5}$$

In this study, however, we have restricted our attention to a simple fourth-order polynomial model, given by

$$\log\ da/dN = k_0 + k_1(\log \Delta K) + k_2(\log \Delta K)^2 + k_3(\log \Delta K)^3 + k_4(\log \Delta K)^4 \tag{6}$$

This equation form does a good job of capturing the sigmoidal shape of the FCGR data (something the log-linear Paris equation cannot do) and has the computational advantage of being a linear regression calculation, as opposed to the more problematic nonlinear regression required for relationships such as Equations (3 and 5). It should be noted, that Equation (6) is not being recommended as a model for general description of FCGR data; it is used here because we have found that it does a satisfactory job of describing fatigue crack growth experimental data. Other expressions have been shown to be superior for recovering crack lengths by integration [e.g., 13], but that was not the goal of this investigation. To illustrate the adequacy of the quartic polynomial for describing *da/dN* versus ΔK data, the average coefficients of determination for each of the four test series are summarized in Table 2. These values indicate that the model is a reasonable reflection of the data in the intervals where the data exist. Whereas the curve fits can be expected to provide good interpolation *between* data points, they clearly cannot be used to extrapolate beyond the domain of the data.

For this analysis, the quartic curve fits were used to estimate the *da/dN* values for each experiment (specimen) at $\Delta K = 2, 4, 6, \ldots 20$ ksi·√inch. Because the starting stress intensity range was 3-5 ksi·√inch in these experiments, the $\Delta K = 2$ ksi·√inch values were not used in the statistical analysis. Recall also, that at the higher ΔK ranges, the data was censored to remove points representing invalid plastic zone sizes according to ASTM E 647. With two experimental data sets (baseline and corroded) established at a given value of ΔK, the issue of whether the data sets can be assumed equivalent, i.e., representative of a single population, can be addressed. The following statistical hypothesis has been posed:

H0: At the $\alpha = 0.01$ significance level, the mean baseline material *da/dN* is equal to the corroded material mean *da/dN* at a given ΔK

To explore this question, an unpaired Student *t*-test was run on each ΔK group. If the null hypothesis was rejected based on the data, we concluded that there was a statistically-significant difference between the baseline and corroded material crack growth rates. In the event that we fail to reject the null hypothesis, we must accept the possibility that there is no difference.

Analysis Results: For purposes of the statistical analysis, the four test series summarized in Table 1 were considered separately; no attempt was made at comparing differences due to stress ratio or relative humidity. In Table 3, the ranges of ΔK are listed where statistically-significant differences in the crack growth rates were observed. Consulting Figures 1-4, we can see that the regions of statistically-significant difference can be easily correlated to differences in the plotted data. Also, comparing Figures 1-4 and Table 3, we see that the tendency for differences in *da/dN* to disappear at higher ΔK levels is confirmed by the statistical analysis. Therefore, based on the original specimen thickness, we conclude that these data show a statistically-significant difference between baseline and corroded material, primarily at low to moderate ΔK levels.

It was noted above that a second reason for developing and implementing the current statistical analysis of FCGR data was to address the validity of modeling corrosion as mechanical damage. Doerfler, et al. [6] originally proposed this simplification without supporting data. Scheuring and Grandt [18], based on a subjective analysis, concluded that, for "light" corrosion, differences in FCGR were accounted for by making an appropriate thickness correction. To explore this issue further, another set of statistical analysis runs were made on the 7075-T6 data. A post fracture microscopic examination of the two shortest-life specimens revealed that, on the fracture surface, the thickness lost to corrosion was between 12% and 16% of the nominal thickness. These thickness reduction figures represent average values on the specimen cross section. More detailed analysis may be able to give more localized thickness data, thus allowing improved estimates of the instantaneous value of ΔK at various crack positions, but it is felt that this added sophistication will not affect the engineering accuracy of the material characterization. Using the 16% thickness reduction figure as representative (and recalling that such values were not available for all the specimens), each specimen thickness was reduced by 16% and modified ΔK's were computed for each data point. The resulting *da/dN* versus ΔK are plotted in Figures 5-8. In the Figures, it can be seen that the corroded material data now lies on top of the baseline data throughout a much larger range of ΔK. In fact, as Table 2 shows, the regions of statistically-significant differences for the reduced thickness case effectively disappear. The exception to this trend is in the $R = 0.05$, < 15% R.H. data set, where the corroded material data are seen to lie below the baseline data. In this case, the thickness correction has resulted in corroded material FCGR's that, as shown in Table 2, are significantly slower than the baseline material. This result is unexpected and is probably not correct physically. Under no circumstances would we expect corrosion damage to retard crack growth. This test series suggests that the thickness correction should be fully applied at higher ΔK levels. In spite of this unusual occurrence, the other test series are found to confirm the assumption that thickness corrections are sufficient to explain FCGR increases in the presence of corrosion damage.

Conclusions: The statistical analysis presented here has been shown to be useful in assessing the equality of fatigue crack growth rate data. It was implemented to determine whether the experimental treatment of corrosion damage was causing significantly increased crack growth rates over a baseline (noncorroded) material. Analysis of experimental data on aluminum alloy 7075-T6 revealed that, using the nominal specimen thicknesses, the corroded material had accelerated crack growth rates, especially at low ΔK levels. At high levels, the baseline and corroded data tended to coincide. When the corroded specimens were corrected for a 16% corrosion-induced thickness loss, the crack growth rates were found to be statistically equivalent over much wider ΔK ranges. This finding represents the first objective evidence that corrosion damage can be modeled as simply a thickness loss.

References:

[1] G.S. Campbell, and R. Lahey, 1984, "A Survey of Serious Aircraft Accidents Involving Fatigue Fracture," *International Journal of Fatigue*, 6(1), 25-30.

[2] J.C.I. Chang, 1995, "Aging Aircraft Science and Technology Issues and Challenge and USAF Aging Aircraft Program," *Structural Integrity in Aging Aircraft*, AD-Vol. 47, American Society of Mechanical Engineers, 1-7.

[3] J.P. Chubb, T.A. Morad, B.S. Hockenhull and J.W. Bristow, 1991a, "The Effect of Exfoliation Corrosion on the Fatigue Behavior of Structural Aluminum Alloys," *Structural Integrity of Aging Airplanes*, Atluri, Sampath and Tong, eds., 87-97.

[4] J.P. Chubb, 1991b, "Effect of Environment and Corrosion on the Fatigue of Structural Aluminum Alloys," *Aluminum Today*, September 1991, 44-49.

[5] J.P. Chubb, T.A. Morad, B.S. Hockenhull and J.W. Bristow, 1995, "The Effect of Exfoliation Corrosion on the Fracture and Fatigue Behavior of 7075-T6 Aluminum," *International Journal of Fatigue*, 17(1), 49-54.

[6] M.T. Doerfler, A.F. Grandt, R.J. Bucci and M. Kulak, "A Fracture Mechanics Based Approach for Quantifying Corrosion Damage," *Proceedings of the Tri-Service Conference on Corrosion*, Orlando Florida, 433-444, 1994.

[7] N.R. Draper and H. Smith, *Applied Regression Analysis*, 2nd edition, John Wiley & Sons, 1981.

[8] H. Ghonem and S. Dore, 1987, "Experimental Study of the Constant-Probability Crack Growth Curves Under Constant Amplitude Loading," *Engineering Fracture Mechanics*, 27(1), 1-25.

[9] W.R. Hendricks, 1991, "The Aloha Airlines Accident - A New Era for Aging Aircraft,", *Structural Integrity of Aging Airplanes*, Atluri, Sampath and Tong, eds., 153-165.

[10] D.W. Hoeppner, L. Grimes, A. Hoeppner, J. Ledesma, T. Mills and A. Shah, 1995, "Corrosion and Fretting as Critical Aviation Safety Issues: Case Studies, Facts and Figures from U.S. Aircraft Accidents and Incidents," *International Committee on Aeronautical Fatigue*, 22 pages.

[11] J. Lincoln, "The USAF Approach to Attaining Structural Integrity of Aging Aircraft," *Structural Integrity of Aging Aircraft*, AD-Vol. 47, American Society of Mechanical Engineers, 9-19.

[12] G. Maymon, 1996, "The Problematic Nature of the Application of Stochastic Crack Growth Models in Engineering Design," *Engineering Fracture Mechanics*, 53(6), 911-916.

[13] M.S. Miller and J.P. Gallagher, "An Analysis of Several Fatigue Crack Growth Rate (FCGR) Descriptions," *Fatigue Crack Growth Rate Measurement and Data Analysis*, ASTM STP 738, S.J. Hudak and R.J. Bucci, eds., 1981, 205-251.

[14] P.C. Paris and F. Erdogan, 1960, "A Critical Analysis of Crack Propagation Laws," *Journal of Basic Engineering*, 85, 528-534.

[15] G.C. Salivar and D.W. Hoeppner, "A Weibull Analysis of Fatigue-Crack Propagation Data From a Nuclear Pressure Vessel Steel," *Engineering Fracture Mechanics*, 12, 1979, 181-184.

[16] A. Saxena, S.J. Hudak and G.M. Jouris, "A Three Component Model for Representing Wide Range Fatigue Crack Growth Data," *Engineering Fracture Mechanics*, 12, 1979, 103-115.

[17] Scheuring, J.N. and A.F. Grandt, 1994, "An Evaluation of Fatigue Properties of Aging Aircraft Materials," USAF Structural Integrity Program Conference, San Antonio TX.

[18] Scheuring, J.N. and A.F. Grandt, 1995, "An Evaluation of Aging Aircraft Material Properties," *Structural Integrity of Aging Aircraft*, AD-Vol. 47, American Society of Mechanical Engineers, 99-110.

[19] Schutz, W., 1995, "Corrosion Fatigue - The Forgotten Factor in Assessing Durability," 15th Plantema Memorial Lecture, 18th Symposium of the International Committee on Aeronautical Fatigue.

[20] D.A. Virkler, B.M. Hillberry and P.K. Goel, 1979, "The Statistical Nature of Fatigue Crack

Propagation," *Journal of Engineering Materials Technology*, 101, 148-153.
[21] J.N. Yang and S.D. Manning, 1996, "A Simple Second Order Approximation for Stochastic Crack Growth Analysis," *Engineering Fracture Mechanics*, 53(5), 677-686.

Table 1: Test Matrix for 7075-T6 Material Showing Number of Specimen Replications

Material Condition	R = 0.05		R = 0.50	
	< 15% R.H.	> 85% R.H.	< 15% R.H.	> 85% R.H.
Baseline	6 specimens	6	6	6
Corroded	4	4	4	4

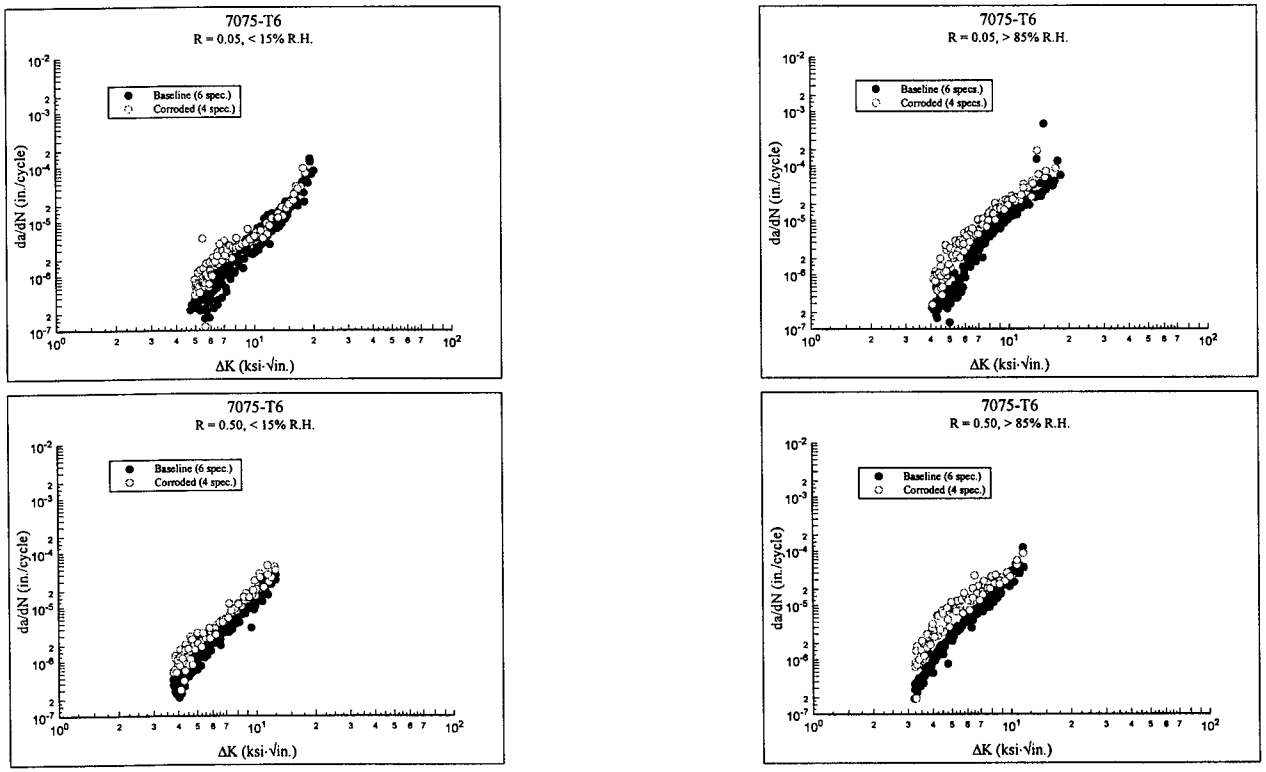
Table 2: Mean Coefficients of Determination From Quartic Curve Fits of *da/dN* vs. ΔK Data

	Mean Coefficient of Determination, r^2		
Test Series	Baseline	No Thickness Corr.	16% Thickness Corr.
R = 0.05, < 15% R.H.	0.9900	0.9355	0.9355
R = 0.05, > 85% R.H.	0.9781	0.9637	0.9637
R = 0.50, < 15% R.H.	0.9772	0.9714	0.9719
R = 0.50, > 85% R.H.	0.9806	0.9620	0.9612

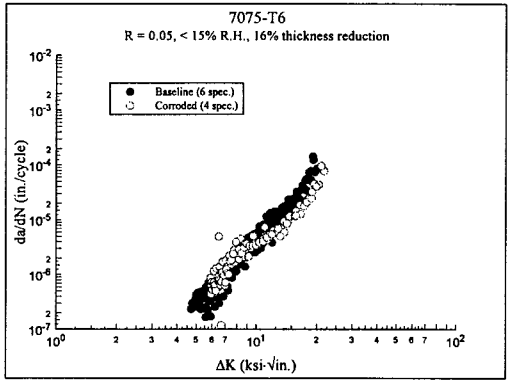
Table 3: Range of Statistically-Significant Differences in *da/dN* vs. ΔK: 7075-T6

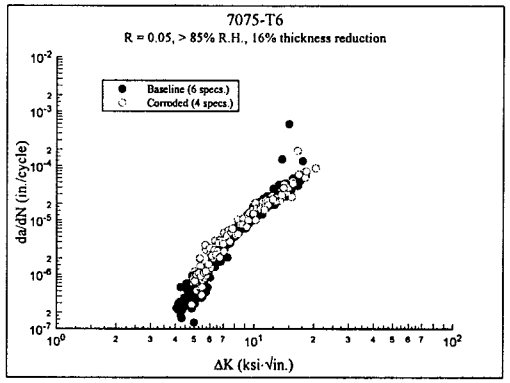
	ΔK Range of Statistically-Significant Difference (ksi·√inch)	
Experiment	No thickness correction	16% thickness correction
R = 0.05, < 15% R.H.	6-8	14-18
R = 0.05, > 85% R.H.	6-12	6*
R = 0.50, < 15% R.H.	4-12	--
R = 0.50, > 85% R.H.	4-8	--

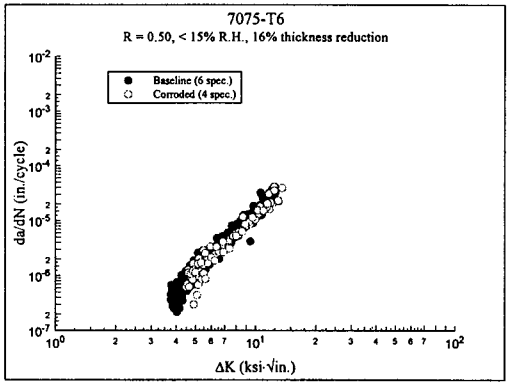
* At ΔK = 6 ksi·√in, significance was found when *t*-test was run with at failed equal variance test (p=0.0465)

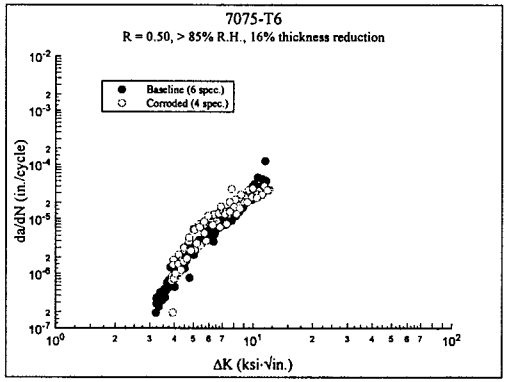


Figures 1-4: Fatigue Crack Growth Rates Based on Nominal Thickness









Figures 5-8: Fatigue Crack Growth Rates Corrected for Thickness Loss

PROBABILISTIC LIFE ASSESSMENTS OF ARBITRARY SHAPED SCARRED BOILER TUBES

K. Zarrabi
School of Mechanical and Manufacturing
The University of New South Wales
Sydney 2052, Australia

H. Zhang
School of Mechanical and Manufacturing
The University of New South Wales
Sydney 2052, Australia

Abstract: This paper describes a practical algorithm and demonstrates its use for probability life analysis of boiler tubes based on the Monte Carlo simulation and Weibull analysis. The tube can have arbitrary shaped scars, gouges, or localised damages with variable rate for loss of its wall thickness. To the authors' knowledge this is for the first time that such an algorithm has been developed. The algorithm has been coded in a computer program coined AUSI-TL1 using an object oriented paradigm.

Key Words: Arbitrary shaped scars; AUSI-TL1; boiler tube; Monte Carlo simulation; probability of failure; Weibull probability analysis.

INTRODUCTION: The need for a reliable boiler tube life assessment has very well been identified in the power, refinery, and chemical processing industries. The function of a fossil-fuelled boiler is to increase the internal (thermal) energy of water and in doing so converts water to superheater steam. The major part of the added internal energy is converted into mechanical work by the steam turbine. The turbine normally runs a generator whose function is to convert the mechanical work into electrical energy. Boiler tubes act as heat exchangers. Hot water or steam runs inside (water or steam side) of a typical boiler tube whereas corrosive and erosive combustion gases flow over its outside surface (fire side). As a result, a boiler tube is continually damaged, gouged and scarred while the tube material is subjected to non-linear creep and plastic deformations. Figure 1 depicts part of a scarred tube schematically. The source of the scarring may include erosion from ash left over from local combustion, thermal oxidation and spalling, steam washes from nearby failed tubes, and nicks and gouges resulting from improper maintenance practices. What ever the source of damage, it is a complex problem that has not been solved precisely to date. One of the limiting factor in assessing a boiler tube life is lack of pertinent loading histories and material properties; we shall talk further about this problem later.

In absence of the precise calculations for the tube life, the plant engineer either replaces the damaged tubes prematurely to avoid tube failures, or tolerates the boiler shut-down because of tube failures. Both of these circumstances cause considerable electricity and financial losses and they are not satisfactory.

Because of its importance to the industrial steam generators, organisations such as Ontario Hydro in Canada, the old CEGB in UK, Babcock & Wilcox, APTECH, and EPRI in USA as well as the authors in Australia have spent considerable research efforts and developed computer programs that provide estimates of a boiler tube life; see for example: Simonen and Jaske [1], Viswanathan, et al. [2], and Zarrabi [3] [4]. Other computer codes that may be mentioned includes: NOTIS by Babcock & Wilcox, TUBECALC by Power House Analytical, and TUBEPRO by Structural Integrity Assoc., all in USA; there are several other proprietary codes. The main limitation of these codes is that they assume the worn or gouged area extends over the entire tube surfaces and ignore the fact that, normally, a tube is scarred or worn in a localised manner. The lack of recognition of the localised nature of the tube damages, usually leads to an overly conservative estimate of a boiler tube life. Also, the above codes normally assume a constant thinning rate for tube wall-thickness. In practice, a tube might be subjected to rapid wall-thickness loss for a period of time. To stop the tube thickness loss, the tube may then be shielded and therefore reduces its thickness loss to almost nil over its remaining life. Therefore, it is important that a computer code for the tube life assessment to allow for a variable thinning rate. To minimise conservatism, the authors have recently developed a computer program coined AUSI-TL1 that estimates the life of a boiler tube containing a scar (Zarrabi and Zhang [5] [6]). The shape and extent of the scar are arbitrary; and this is for the first time to the authors' knowledge that such a computer code has been developed. The other advantage of the AUSI-TL1 code over other existing codes is that AUSI-TL1 considers the tube thickness loss in a precise-wise linear manner so that a variable tube thinning rate may be accounted for.

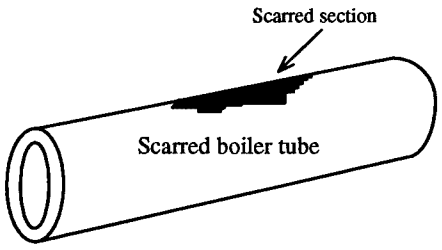


Figure 1 Part of A Boiler Tube Containing A Scar

The above computer codes (including the AUSI-TL1) for tube life assessments are, however, based on deterministic approaches that produce a single value of the tube life assuming that input data are precise and not contaminated with uncertainties. The reality is, however, different. In practice, there is a degree of uncertainty in each of the input values such as: steam pressure, metal temperature, material properties, etc.. At best, we may estimate a range that is likely to contain the right value of the concerned input variable. Therefore, the computation of a single value of the tube life may become misleading. Instead a probabilistic approach may be more appropriate.

This paper describes the extension of the AUSI-TL1 code to include a probabilistic life assessment module. Although there are some computer codes that use a probabilistic method for tube life assessments (see, for example, Sprung and Zilberstein [7]), our probabilistic method, for the first time, considers localised arbitrary tube scars with variable thinning rate.

PROBABILISTIC APPROACH: Our probabilistic formulation for tube life assessments combines the Monte Carlo simulation and the Weibull analysis. It is also based on the same equations as our deterministic formulation (for description of our deterministic formulation see, Zarrabi and Zhang [5] [6]). Therefore, our probabilistic method accounts for localised scars and gouges with variable thinning rate similar to our previously developed deterministic method. The algorithm of our probabilistic method consists of the following main steps:

1. Input Data: The input data, X_i, for the tube life calculations are: loading data (such as steam pressure, metal temperature), data related to tube geometry (such as thickness, radius, scarred dimensions), and material properties (such as yield strength, ultimate tensile strength, creep rupture data). These data normally contain a degree of fluctuation, scatter, and uncertainty. If the mean value for each input data is X_{Mi} and its stochastic range is δX_i, then:

$$X_i = X_{Mi} + \delta X_i / 2 \qquad (1)$$

The essence of the Monte Carlo simulation is in computing a tube life for randomly selected X_i in such a way that equation 1 is satisfied. This is accomplished by using a probability density function (PDF) for each input data over its specified range.

2. Probability Density Function (PDF): In principle, the PDF of each input data may be obtained by a statistical analysis of available records from plants. Provided the archived data from plants are available and reliable, it is likely that such an analysis provides different PDF for different input data and different boilers. To devise a generic algorithm

that would be applicable to a wide range of boilers and as a first step, we have assumed a uniform distribution for PDF, ie:

$$\text{PDF}(X_i) = 1/\delta X_i \tag{2}$$

Equation 2 gives equal weight to all possible random values that fall within each range of input data defined by equation 1. However, in reality, it may be that the probability of values close to the mean values to be more than other random values. For such cases, a Gaussian PDF may be more appropriate to adopt. The study of various distributions for PDF and their effects on the tube life will be the subject of future publications.

Using the uniformly distributed PDF defined by equation 2 and a library function available in BORLAND C^{++} compiler, we have computed random values for each input data.

3. Multiple Life Computation: Let t_M to represent the tube life computed on the basis of the mean values of each input data and t_i (where $i = 1,2,3,\ldots,N$ with N is the number of randomly sampled values of each input data) to represent those obtained on the basis of random values of each input data. As mentioned above, t_M and t_i are computed using our previously developed deterministic computer code: AUSI-TL1. In this way N random tube lives are computed.

Our objective is to estimate the probability of t_M. To this end, let define a performance function, G, where:

$$G = t_i - t_M \tag{3}$$

The probability of t_M increases whenever t_i becomes equal to t_M where $G = 0$. In another words, $G = 0$ defines a probability failure surface where each point on this surface provides a probable tube failure point.

4. Weibull Probability Analysis: The first step in performing a Weibull analysis is to rank or rearrange the randomly computed tube lives (t_i) and store them in a vector (t_j) so that the shortest tube life is the first element of the vector and the longest life is the last element in the vector. Next the median rank (F_j) of t_j are computed using (Johnson [8]):

$$F_j = 100\ (j - 0.3)/(N + 0.4) \tag{4}$$

where j is the order number of tube lives in the vector t_j. Then we assume either two parameter (equation 5) or three parameter (equation 6) Weibull probability distributions, ie:

$$F(t) = 1 - e^{-(t/\eta)^{\beta}} \tag{5}$$

where η and β are the two parameters that are determined by fitting equation 5 to F_j and t_j data, or:

$$F(t) = 1 - e^{-((t-t_0)/\eta)^{\beta}} \tag{6}$$

where η, β, and t_0 are the three parameters that are determined by fitting equation 6 to F_j and t_j data. $F(t)$ in equations 5 and 6 represents the probability of any life t. The Weibull probability distribution that more closely fit the random life data will be selected as final probability distribution. Finally, the probability of t_M (ie, $F(t_M)$) is computed by using $t = t_M$ in either equation 5 or equation 6.

The computer program, AUSI-TL1, has been extended to include an object that performs tube life probability calculations that is based on the above algorithm.

EXAMPLES OF PROBABILITY COMPUTATIONS OF TUBE LIFE USING AUSI-TL1 CODE: One example is considered here that illustrates the application of the algorithm outlined above. The main input data with their estimated uncertainties are as follows:

Material:	SA213-T22	
Time in service:	60,000±100	Hours
Steam internal pressure:	20±4 MPa	
Metal temperature:	590±5 °C	
Outside diameter:	50±2.5	mm
Thickness:	9±1.0 mm	
Scar length:	5±1.5 mm	
Scar width:	3±1.5 mm	
Scar depth:	2±1.0 mm	

The above data are selected in such a way that creep deformation becomes a significant event for this example. Setting $N = 100$, the predicted tube lives by AUSI-TL1 code using the above input data are summarised as a histogram in Figure 2. The cumulative probability failure is plotted in Figure 3. The closeness of the plotted points to a linear distribution in Figure 3 indicates that the two-parameter Weibull provides a good representation of the probability distribution function. The outputs of the Weibull analysis are summarised in Table I.

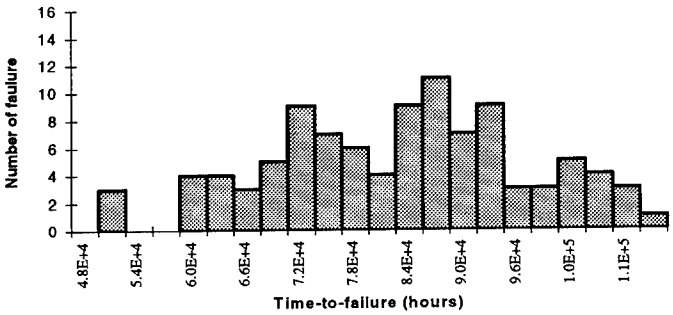


Figure 2 Histogram of Boiler Tube Lives

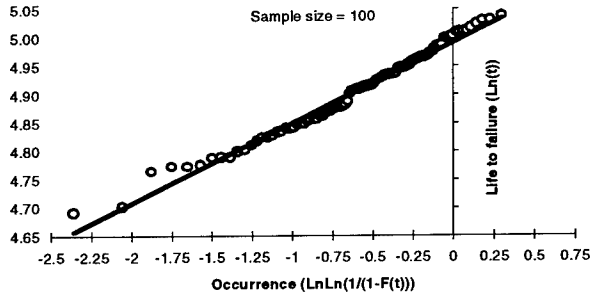


Figure 3 Cumulative Probability of Failure

Table I Summary of Results From Weibull Analysis

Failure Mode	Creep Rupture
Mean Life	82600
Probability of Failure	61.2%
β	7.01023
η	86740.2

The Weibull probability density function is shown in Figure 4. The plot in Figure 4 can be used to obtain the probability of failure for any given tube life. With pass of time, the risk of tube failure will increase. The trend of the probability of tube failure versus time is shown in Figure 5. Figure 5 indicates that for less than 60,000 hours time-in-service, the

risk of tube failure is small. The risk of tube failure increases rapidly for time between 60,000 hours and 100,000 hours. The risk of failure beyond 100,000 hours remains virtually unchanged. Figures 5 to 9 shows the effect of various input data on the probability of failure for the tube.

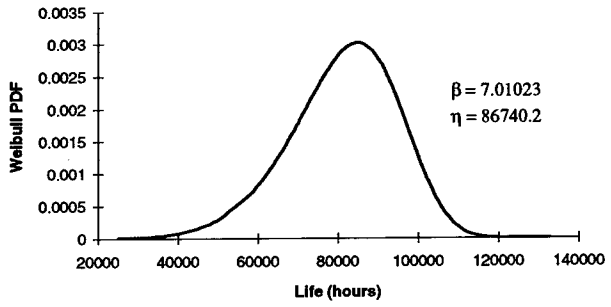


Figure 4 Weibull Probability Density Function

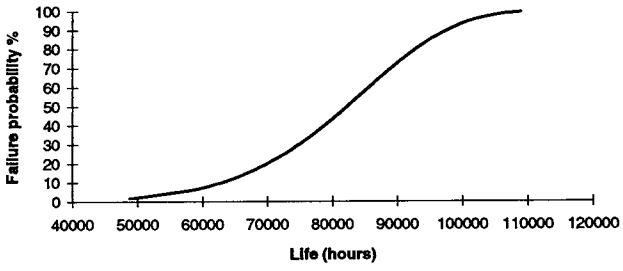


Figure 5 Failure Probability vs. Life

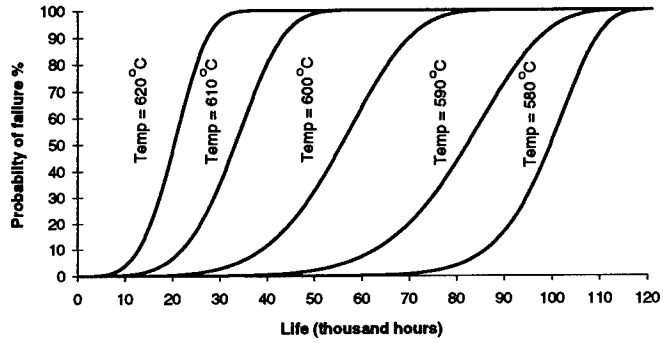


Figure 6 Effects of Tube Temperature on Failure Probability

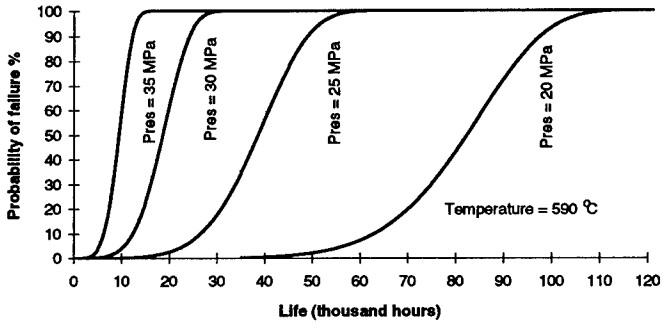


Figure 7 Effects of Stream Pressure on Failure Probability

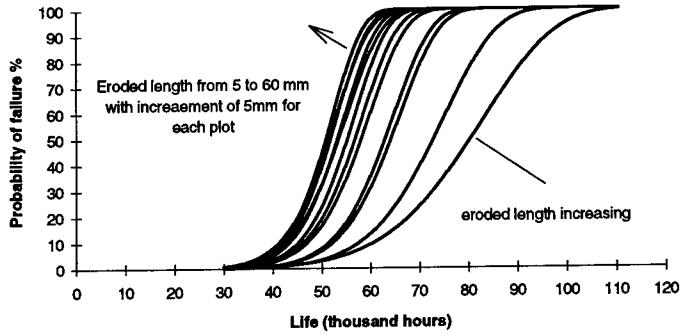


Figure 8 Effects of Scarred Length on Failure Probability

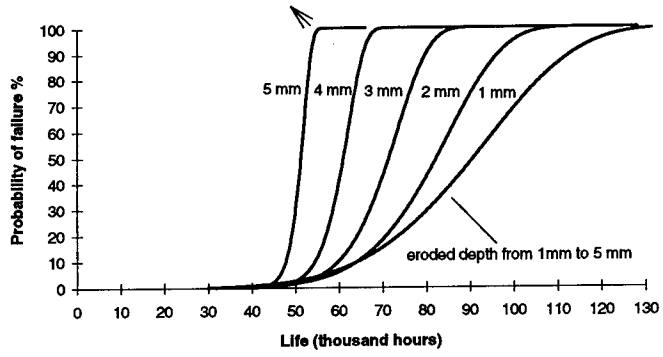


Figure 9 Effects of Scarred Depth (or Thinning) on Failure Probability

CONCLUDING REMARKS: This work shows that a combination of the Monte Carlo simulation and Weibull analysis provides a practical algorithm for estimating the probability of a failure for boiler tubes containing localised scars and damages and subjected to a variable rate of wall-thickness loss. The validity of results depends on how closely the randomly computed tube lives match the two-parameters or three-parameters Weibull probability distribution functions. The above probability algorithm for tube life assessments has been coded into the AUSI-TL1 computer code for ease of application.

There may be circumstances for which the Weibull probability distribution functions do not provide good representations. For such cases, it may be that the direct Monte Carlo technique provides a more accurate estimates for the probability of tube failure. Currently, we are extending the AUSI-TL1 code to include an object for probability analysis based on the direct Monte Carlo analysis. Having developed this object, we should be in a position to compare the Weibull and direct Monte Carlo algorithms and report the results in future publications.

REFERENCES: [1] Simonen, F. A. and Jaske, C. E.,1985, A Computational Model For Predicting The Life of Tubes Used in Petrochemical Heater Services, Trans. ASME, Journal of Pressure Vessel Technology, Vol. 107.
[2] Viswanathan, R., Paterson, S. R., Grunloh, H. and Gehl, S., 1994, Life Assessment of Superheater/Reheater Tubes in Fossil Boilers, Trans. ASME, Journal of Pressure Vessel Technology, Vol. 116.
[3] Zarrabi, K., 1992, Estimation of Boiler Tube Life in Presence of Corrosion and Erosion Processes, Int. J. Pres. Ves. & Piping, Vol. 53, pp351-358.
[4] Zarrabi, K., 1993, A Computational Code For Obtaining Creep and Plastic Collapse Lives of Boiler Tubes That Are Subjected to Erosion/Corrosion, Proceedings of The ASME Pressure Vessels and Piping Division Conference, Denver, Colorado, USA, Plant Systems/Components Aging Management, PVP-Vol. 252, pp171-175.
[5] Zarrabi, K. and Zhang, H., Boiler Tube Life Assessments With Localised Erosion/Corrosion, 1995, Proceedings of The Australia Materials Conference, Perth, Australia, Book Two, pp204-209.
[6] Zarrabi, K. and Zhang, H., 1995, Primary Stress in Scarred Boiler Tubes, Int. J. Pres. Ves. & Piping, Vol. 65.
[7] Sprung, I. and Zilberstein, V. A., 1992, Superheater and Reheater Boiler Tube-life Estimation, Materials Evaluation, pp986-990.
[8] Johnson, L. G., 1964, The Statistical Treatment of Fatigue Experiments, Elsevier Publishing Company, London, England.

ANALYSIS OF ELECTROMECHANICAL SYSTEMS RELIABILITY USING MARKOV MODELS

J. Lu, J.D. Baldwin

School of Aerospace & Mechanical Engineering
University of Oklahoma
Norman, Oklahoma 73019

Abstract: A computational algorithm is presented to estimate the reliability of an engineering system consisting of both electrical and mechanical components. The mechanical component is modeled with a strain-based fatigue reliability model that allows the hazard rate functions of the cycles to failure distribution to be predicted. Any electrical component is assumed to exhibit a constant hazard rate function. Markov state transition probabilities are derived in order to include time varying hazard rate functions in the analysis. The time-dependent reliability of the system is computed using a Markov chain analysis under the assumption that the fatigue-based hazard rate function of the mechanical elements are piecewise constant. An example is given that demonstrates the piecewise constant hazard rate function method applied to an engineering system.

Key Words: Markov chain; Non-stationary; Reliability analysis; System reliability; Variable hazard rate function

Introduction and Background: Evaluation of the reliability of an engineering system is a critical design task. One of the important techniques used to evaluate system reliability is the Markov chain approach. In this technique, a system is assumed to pass from its operational state to any number of failed states, often passing through intermediate states representing the condition of subsystems. The vital role of this approach is to make reliability evaluation of repairable systems feasible by allowing the repair process (i.e., transition from failed to operational states) to be included in the reliability model. Typically, this method is used under the assumption of time-invariant probabilities of transition between the states of the system, a situation commonly encountered with electrical and electronic devices.

Many modern engineering systems, however, are combinations of electrical and mechanical components. Here, mechanical components are taken to be stressed solids under cyclic loading; the electrical devices can be transistors, integrated circuits, coil windings and so on. These systems must be considered non-stationary in the Markov sense because the failure rates of most mechanical devices are not constant, typically due to wear-out processes such as metal fatigue. In that case the reliability would take the general form

$$R(t) = \exp\left(-\int_0^t \lambda(t')dt' \right) \tag{1}$$

The calculation of the evolution of these non-stationary Markov models, however, is complicated and is difficult to do by classical methods. In addition, even for a large system with stationary Markov properties, the process of calculation becomes very cumbersome.

The use of Markov models in reliability analysis is familiar in electrical systems [10] and several authors have addressed the analytical solution of non-stationary Markov models in the context of arbitrary repair rate functions [3, 14] and mean time to system failure [15]. A Monte Carlo-based algorithm was used by Lewis and Zhuguo [11] to model the evolution of nonstationary models with explicit expressions of the time varying hazard rate function. It is also interesting to note that, although they are not directly applicable to the combined analysis given here, the probabilistic cumulative damage models of Bogdanoff and Kozin [2] use Markov chains to model the damage evolution within a damaging material. Unlike electrical systems, most mechanical devices or systems exhibit hazard rate functions that vary in time. It is reasonable to expect that failure modes that depend on time in service are the result of wear-out phenomena and will not exhibit constant hazard rates. For many mechanical components and devices, fatigue is a common cause of failure and the literature [7,16,17] shows that fatigue life has been modeled with both the Weibull and lognormal distributions. Recently, Baldwin and Thacker [1] developed a strain-based fatigue reliability analysis method which can be used to estimate the fatigue hazard rates at a given number of load cycles.

The focus of this paper is to report a method of implementing a combined mechanical-electrical reliability analysis and to evaluate system reliability using the resulting non-stationary Markov models. This approach will be referred to as the piecewise-constant hazard rate (PHR) method. Here we summarize the details of the PHR method which can be used to evaluate the reliability of general mechanical-electrical systems.

Markov Chain Analysis: In this paper, we consider only discrete state-discrete time Markov chains to model the random behavior of an engineering system consisting of both electrical and mechanical components. Mathematically, a continuous time non-stationary Markov process can be formulated as a first-order, linear ordinary differential equation with variable coefficients. If the ODE is discretized in time, the resulting difference equation describing the discrete time process's evolution is of the form

$$\boldsymbol{P}(t_m) = \boldsymbol{P}(t_{m-1})\boldsymbol{A}(t_m) \tag{2}$$

where $\boldsymbol{P}(t_m)$ is the state probability vector

$$\boldsymbol{P}(t_m) = \left[P_1(t_m) \;\; P_2(t_m) \;\; . \;\; . \;\; . \;\; P_N(t_m)\right] \tag{3}$$

$A(t)$ is the state transition probability and t_m (= $m\,\Delta t$; m = 1, 2, ...) is the time elapsed during the analysis. It is convenient to express the state transition probabilities in matrix form as

$$\boldsymbol{A}(t_m) = \begin{pmatrix} A_{11}(t_m) & A_{12}(t_m) & \cdots & A_{1N}(t_m) \\ A_{21}(t_m) & A_{22}(t_m) & \cdots & A_{2N}(t_m) \\ \vdots & \vdots & \ddots & \vdots \\ A_{N1}(t_m) & A_{N2}(t_m) & \cdots & A_{NN}(t_m) \end{pmatrix} \tag{4}$$

This transition probability matrix contains all the available information about the passage of the system among its states at the m-th time step. Because $A_{ij}(t_m)$ is a conditional probability based on the requirement that the system is in state i at time t_{m-1}, the system at time t_m may either transfer to any of the other N-1 states, or remain in state i. Elements of the i-th column of $\mathbf{A}(t_m)$ represent transitions *to* state i from the other states; similarly, elements of the i-th row describe transitions *from* state i to the other states. If $i \neq j$, $A_{ij}(t_m)$ is the probability that the system leaves state i and transfers to state j at the m-th time step, given that the system operates in state i through the (m-1)-th time interval. If the system leaves state i due to the failure of the corresponding subsystem, the

probability of that transition can be stated in terms of the subsystem's reliability function as

$$A_{ij}(t_m) = 1 - \frac{R(t_m)}{R(t_{m-1})} \tag{5}$$

If we consider the time discretization $t = t_m = m\,\Delta t$, and assume that Δt is small enough to make negligible, the state transition probability, Equation (5), becomes

$$A_{ij}(t_m) \approx \int_{t_{m-1}}^{t_m} \lambda_{ij}(t')dt' \tag{6}$$

Equation (6) is useful in that the hazard function $\lambda_{ij}(t)$ has not been assumed constant [12]. If the system leaves state i due to the repair of one of its subsystems, maintainability $G(t)$ can be stated as the probability that a failed subsystem is restored to one of its operating states at time t. The repair transition probability from state i to state j is

$$A_{ij}(t_m) = \frac{G_{ij}(t_m) - G_{ij}(t_{m-1})}{1 - G_{ij}(t_{m-1})} \tag{7}$$

and in a manner similar to the reliability function, the discrete time state transition probability can be written in terms of the repair rate function as

$$A_{ij}(t_m) \approx \int_{t_{m-1}}^{t_m} \mu_{ij}(t')dt' \tag{8}$$

For the probability of the system remaining in state i (i.e., $i = j$),

$$A_{ii}(t_m) = 1 - \sum_{j=1}^{N} \int_{t_{m-1}}^{t_m} \left(\mu_{ij}(t') + \lambda_{ij}(t')\right)dt' \tag{9}$$

Note that by definition, for a given state transition, either μ_{ij} or λ_{ij} (or both) must be zero; the transition can be due either to failure or to repair, but not both.

It can be seen that the transition probability matrix, Equation (4), is complicated by the integrations in Equations (6), (8) and (9). By the PHR method, we assume $\lambda_{ij}(t)$ and $\mu_{ij}(t)$ are constant within each time interval $((m-1)\Delta t, m\Delta t)$ and are denoted $\overline{\lambda}_{ij}(t_m)$ and $\overline{\mu}_{ij}(t_m)$. The piecewise constant hazard (failure) rate $\overline{\lambda}_{ij}(t_m)$ and repair rate $\overline{\mu}_{ij}(t_m)$ functions are given to a first approximation by

$$\overline{\lambda}_{ij}(t_m) = \frac{1}{2}\left(\lambda_{ij}(t_{m-1}) + \lambda_{ij}(t_m)\right) \tag{10}$$

$$\overline{\mu}_{ij}(t_m) = \frac{1}{2}\left(\mu_{ij}(t_{m-1}) + \mu_{ij}(t_m)\right) \tag{11}$$

The transition probability matrix, Equation (4), is thus established using Equations (10-11).

For a non-stationary discrete state, discrete time Markov model, the time period of the analysis is decomposed into time steps and the transition probabilities are allowed to vary step by step. In the general case of a non-stationary Markov model, the transitional probabilities are time-dependent

within each time interval. The PHR method arises by assuming constant transition probabilities for the system within each time interval and combining the time-varying transition matrices to calculate the state probability of the system at a given time. The general equations for the discrete time Markov model can be directly used to solve these non-stationary problems.

For the system consisting of N states, the probabilities of the system being in each state at time t_m can be expressed in matrix form as given in Equations (2-4). For $t_{m-1} \geq 0$, the state probability vector recursion is summarized by

$$\boldsymbol{P}(t_m) = \prod_{i=1}^{m} \boldsymbol{P}(t_0)\, \boldsymbol{A}(t_i) \tag{12}$$

Equation (12) [9] is a general equation to evaluate the probability of being in each state in the system at any time step. If we know the initial state of the system $\boldsymbol{P}(t_0)$ and the state transition matrix $\boldsymbol{A}(t)$, we can compute the probability that the system will be in any given state at any time t_m. Because the state transition matrix varies at each time step (due to the piecewise-constant transition probabilities), we therefore have an algorithm for estimating the non-stationary Markovian system. The reliability of the system at the m-th time interval is the probability of the system being in any of its operational states. For an N-state system, the reliability is given by

$$R(t_m) = \sum_{\text{operational states}} P(t_m) = 1 - \sum_{\text{failed states}} P(t_m) \tag{13}$$

Probabilistic Fatigue Analysis: The fatigue reliability model used here is described in detail by Baldwin and Thacker [1]; only a brief summary of the details will be given here. In the strain-based fatigue model, the basic assumption involved is that the appearance of a crack in a component can be related to the fracture of a small specimen in a laboratory fatigue test.

It is known that the lives to failure of smooth test specimens tested at given strain amplitude can exhibit considerable scatter [4,5,6,8,13]. The three-parameter Weibull distribution will be used to model the scatter observed in fatigue tests. The probability density function for the Weibull distribution can be expressed as

$$f\left(N_f(\Delta\epsilon)\right) = \left(\frac{\beta}{N_c - N_m}\right)\left[\frac{N_f(\Delta\epsilon) - N_m}{N_c - N_m}\right]^{\beta-1} \exp\left[-\left(\frac{N_f(\Delta\epsilon) - N_m}{N_c - N_m}\right)^{\beta}\right] \tag{14}$$

where β is the Weibull distribution shape parameter, N_c is the Weibull distribution characteristic life and N_m is the Weibull distribution minimum life. $N_f(\Delta\epsilon)$ is the number of cycles to failure at a given strain amplitude $\Delta\epsilon$, given implicitly in the strain-life relationship

$$\frac{\Delta\epsilon}{2} = \frac{\sigma_f'}{E}(2N_f)^b + \epsilon_f'(2N_f)^c \tag{15}$$

The cycle ratio η

$$\eta = \frac{N_m}{N_c - N_m} \tag{16}$$

is assumed to be a material constant based on the data analysis of Tanaka, et al. [16]. By the assumption that the strain-life equation is taken to model median failure behavior, the parameters N_m, N_c are then given by

$$N_m = \frac{\eta N_f(\Delta\epsilon)}{[-(\ln 0.5)]^{1/\beta}+\eta} \tag{17}$$

$$N_c = \frac{(1+\eta) N_f(\Delta\epsilon)}{[-(\ln 0.5)]^{1/\beta}+\eta} \tag{18}$$

The hazard rate function of the component under fatigue load at any time (number of cycles) is given by

$$\lambda(N) = \frac{R(N_1) - R(N_2)}{R(N)(N_2 - N_1)} \tag{19}$$

where N_1 and N_2 define a narrow interval containing N.

Combined Electromechanical Reliability Analysis: To illustrate the combined electromechanical system reliability (including fatigue), we now consider a system consisting of a electrical unit E and a mechanical unit M shown schematically in Figure 4. Such a system could be realized in many different ways, for example a pump-motor set (where the motor is modeled as a purely electrical device). The technique is also applicable to systems where the electrical and mechanical components are both integral to the operation of the system, but are otherwise unrelated. These two units are assumed to act in series, i.e., if either unit fails to operate, the system fails to operate. The mechanical unit M is experiencing a load history as shown in Figure 5; this loading block is assumed to repeat every 1 hour. The material of the mechanical unit is a carbon steel with strain-life parameters as given in Table 1. A stress concentration $K_t = 2.3$ is assumed. The fatigue behavior of the material is assumed to be such that the strain-life Weibull distribution shape parameter is $\beta = 2.0$, and the cycle ratio is $\eta = 0.2$. No repair is available for this unit, i.e., $\mu_M = 0$. The constant failure rate of the electrical unit E is $\lambda_E = 10^{-5}\text{hr}^{-1}$. A failed electrical unit has a repair rate $\mu_E = 0.1$ hr^{-1}.

The state space of this system can be summarized as shown in Table 2. Figure 6 shows the system state space and possible state transitions. Using Equation (4), we can write the state transition matrix as

$$A(t_m) = \begin{pmatrix} 1 - (\lambda_E(t_m) + \overline{\lambda_M}(t_m))\Delta t & \lambda_E(t_m)\Delta t & \overline{\lambda_M}(t_m)\Delta t \\ \mu_E(t_m)\Delta t & 1 - \mu_E(t_m)\Delta t & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The initial condition of the system is assumed to be $P(t_0) = (1\ 0\ 0)$. The reliability of this system is defined as the probability that the system is in this state (state 1), i.e, $R = P_1(t)$.

The combined electromechanical reliability prediction algorithm detailed here has been used to analyze this system. Figure 7 shows that, as expected, the fatigue hazard rate function increases as a function of mission time. This functional characteristic follows directly from the fatigue reliability model and is felt to be representative of the behavior of a population of deteriorating components. As with any Weibull-based hazard rate function, modifying β will affect the shape of

the $\lambda(t)$ curve. Figure 4 also shows the system reliability as a function of mission time for $\Delta t = 6.25$ hours (the suitably short time step that satisfies the Markov assumptions). Also, different values of the strain-life distribution parameters β and η were used in the analysis to observe how the system reliability behaves for variations in these parameters. Figures 5 and 6 reveal that uncertainty in the shape parameter β of the strain-life Weibull distribution can lead to much more uncertainty in the value of the system reliability than does uncertainty in the cycle ratio parameter η. Of course, if we had allowed repair of the mechanical component, state 3 would no longer be a terminal state. Also, changing any of the analysis parameter values would modify the solution.

Conclusions: Today, many modern engineering systems are combinations of electrical and mechanical components. The mechanical components complicate the analysis because of their varying state transition probabilities. These arise from wear-out processes, such as fatigue, that cause temporally-increasing hazard rate functions for the mechanical failure. In this paper, non-stationary, discrete time, discrete state Markov modeling techniques associated with a piecewise-constant hazard rate method were presented to evaluate reliabilities of electromechanical systems. The solution of a sample case illustrates how an appropriately formulated fatigue reliability theory and Markov chain modeling techniques can be used to evaluate the reliability of electromechanical systems simply and efficiently. The Markov models we discussed here are discrete time models, although most engineering systems evolve continuously in time. However, this work shows that a continuous time Markov model can be converted to the discrete time model with the precision of the solution improving as the time step Δt is reduced.

References:

[1] J.D. Baldwin and J. G. Thacker, "A Strain-Based Fatigue Reliability Analysis Method," *Journal of Mechanical Design*, Vol. 117, Number 2(a), 1995, pp. 229-234.

[2] J.L. Bogdanoff and F. Kozin, *Probabilistic Models of Cumulative Damage*, John Wiley & Sons, 1985.

[3] M.H. Branson and B. Shah, "Reliability Analysis of Systems Comprised of Units with Arbitrary Repair-Time Distributions," *IEEE Trans. Reliability*, Vol. R-20, No. 4, 1971, pp. 217-223.

[4] H.T. Corten, G. M. Sinclair and T. J. Dolan, "An Experimental Study of the Influence of Fluctuating Stress Amplitude on Fatigue Life of 75S-T6 Aluminum," *Proceedings of the ASTM*, Vol. 54, 1954, pp. 737-754.

[5] T.J. Dolan, F. E. Richart Jr. and C. E. Work, "The Influence of Fluctuations in Stress Amplitude on the Fatigue of Metals," *Proceedings of the ASTM*, Vol. 49, 1949, pp. 646-682.

[6] A.M. Freudenthal and E. J. Gumbel, "Physical and Statistical Aspects of Fatigue," *Advances in Applied Mechanics*, Vol. 4, H.L. Dryden and T. Von Karman, eds., Academic Press, 1956, pp. 117-158.

[7] E.J. Gumbel, "Parameters in the Distribution of Fatigue Life," *Journal of the Engineering Mechanics Division*, Vol.89, Number EM5, 1963, pp. 45-63.

[8] W.E. Hering and C. W. Gadd, *Experimental Study of Fatigue Life Variability*, Research Publication GMR-555, General Motors Corporation, 1966.

[9] D.L. Isaacson and R. W. Madsen, *Markov Chains: Theory and Applications,* John Wiley & Sons, Inc., 1976.

[10] B.W. Johnson, *Design and Analysis of Fault-Tolerant Digital Systems*, Addison-Wesley, 1989.

[11] E.E. Lewis and T. Zhuguo, "Monte Carlo Reliability Modeling by Inhomogeneous Markov Processes," *Reliability Engineering*, Vol. 16, 1986, pp. 277-296.

[12] N.J. McCormick, *Reliability and Risk Analysis*, Academic Press, Inc., 1981.

[13] G.M. Sinclair and T. J. Dolan, "Effect of Stress Amplitude on Statistical Variability in Fatigue Life on 75S-T6 Aluminum Alloy," *Proceeding of the American Society of Mechanical Engineers*, Vol.75, 1953, pp. 867-872.

[14] C. Singh, R. Billinton, S. Y. Lee, "The Method of Stages for Non-Markov Models," *IEEE*

Trans. Reliability, Vol. R-26, No.2, June 1977, pp. 135-137.
[15] C. Singh, "Equivalent Rate Approach to Semi-Markov Processes," *IEEE Trans. Reliability*, Vol. R-29, No.3, Aug. 1980, pp. 273-274.
[16] T. Tanaka, T. Sakai, and T. Iwaya, "Distribution Characteristics of Fatigue Lives and Fatigue Strengths of Ferrous Metals by the Analysis of P-N Data in the JSMS Data Base on Fatigue Strength of Metallic Materials," *Statistical Research on Fatigue and Fracture*, T. Tanaka, S. Nishijima and M. Ichikawa, eds. Elsevier Applied Science Publishers, 1987, pp. 125-157.
[17] P.H. Wirsching, *The Application of Probabilistic Design Theory to High Temperature Low Cycle Fatigue,* NASA CR-165488, 1981.

Table 1: Strain-life fatigue properties [SAE J1099]

$\sigma'_f = 78{,}000$ psi	$b = -0.073$
$\epsilon'_f = 0.110$	$c = -0.410$
$K' = 71{,}000$ psi	$n' = 0.110$
$E = 29 \times 10^6$ psi	

Table 2: State space of the two component electromechanical system

State	Units Operating	Units Failed	System Condition
1	E, M	---	Operating
2	M	E	Failed (repairable)
3	E	M	Failed (not repairable)

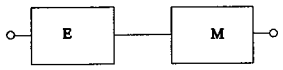


Figure 1: A two component electromechanical system

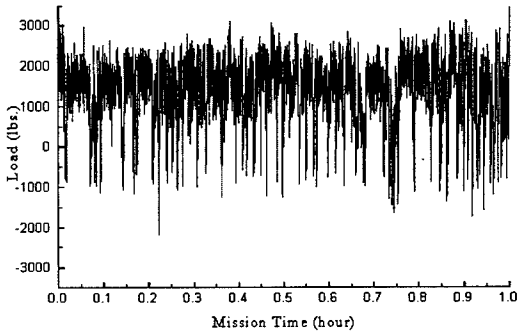


Figure 2: Applied load time history

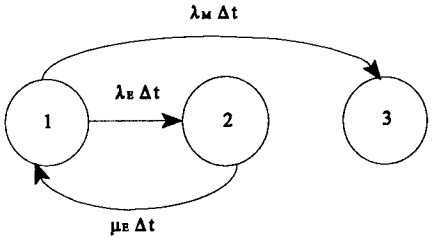
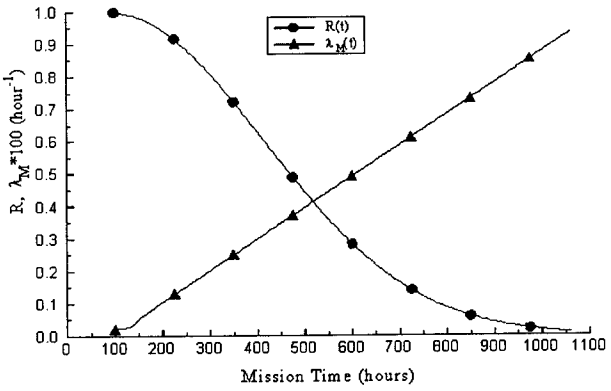


Figure 3: State space diagram for a two component electromechanical system



Fatigue 4: System reliability and fatigue failure rate (unit M)

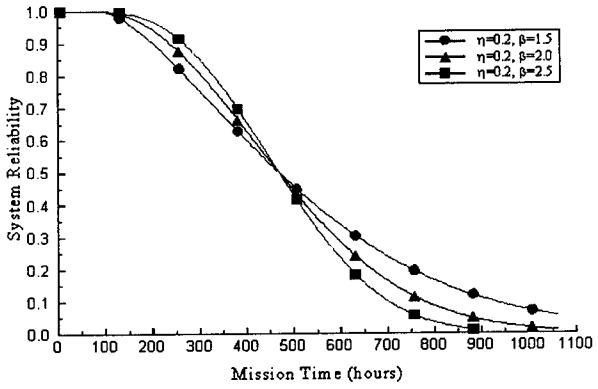


Figure 5: System reliability function - Variable β

A NEW QUANTIFICATION OF THE RELIABILITY OF COMPONENTS SUBJECTED TO ALL TYPES OF FATIGUE LOADS

Dimitri B. Kececioglu and Wendai Wang
Aerospace and Mechanical Engineering Department
The University of Arizona
Tucson, AZ 85721

Abstract: Based on a statistical model of fatigue data, or the P-S-N curves, the distributions of fatigue life are provided in this paper under various conditions. For the situation of constant-amplitude stress with uncertainty of magnitude, the theoretical distribution of fatigue life is derived by using the Theorem of Total Probability and the concept of conditional reliability. For the wide-band stress case, a cycle-counting method is proposed. After deriving the distributions of the magnitude and the mean value of the stress cycles according to the proposed cycle-counting method, a closed form expression for the mean value of the fatigue life is formulated for wide-band stress spectra. The effect of a non-zero mean of stress history is taken into account in this approach. The derived results are illustrated with a numerical example.

Key Words: Fatigue life; reliability analysis; wide-band stress

INTRODUCTION: Metal fatigue is one of the most important failure modes to be considered in mechanical and structural design. The only satisfactory way to prevent fatigue failures during service life is by proper design. Unfortunately, many of the design factors associated with fatigue are subject to considerable uncertainties. However, designers have to make decisions at the design stage. Thus, reliability or probabilistic design methods are necessary and appropriate for fatigue design and analysis. For engineering applications, designers need a simple and reasonably accurate design code to predict the reliability for a specified service life. The approaches of this paper provide a methodology to determine the distribution of the fatigue life (cycles-to-failure or times-to-failure) under various situations.

A STATISTICAL MODEL FOR FATIGUE TEST DATA: The conventional fatigue theories have focused on the S-N diagram, which is based on standard constant-amplitude-stress fatigue tests at several given stress levels. The S-N diagram relates the cycles to failure, N, to the cyclic stress amplitude, S. In many cases, the S-N data plots have a linear trend on log-log scales; i.e.,

$$N S^m = A, \tag{1}$$

or

$$\log N + m \log S = \log A. \tag{2}$$

Because fatigue data have large statistical scatter, the actual fatigue characteristic should not be represented by a single S-N curve but by a family of distributions; i.e., the so-called P-S-N diagram. Hence, the cycles to failure, N, at any stress level, S, is considered to be a random variable. Equation (1) or (2) actually corresponds to the mean or the median curve of the P-S-N diagram [1]. Many articles and books have been published on this topic of statistical analysis of fatigue data. Among those widely used approaches, a simple model,

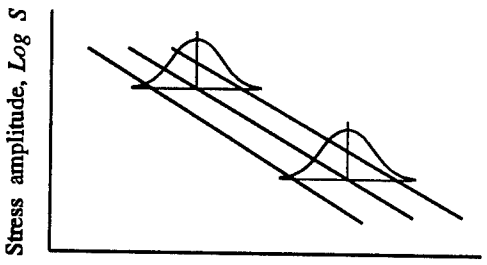


Fig. 1 – The *P-S-N* curves.

pursued by Wirsching and Hsien [2], is that, at any given stress amplitude, S, (1) the fatigue strength exponent, m, is assumed as a non-random constant; and (2) the fatigue strength coefficient, A, is treated as a lognormal random variable.

The value of constant, m, and the estimates of the parameters of variable, A, (the mean and standard deviation of $\log A$; i.e., $\bar{A}'$ and $\sigma_{A'}$, or the coefficient of variation of A; i.e., C_A) can be obtained by employing the least-squares or the MLE method [3] to testing data. All approaches in this paper are based on this model.

From Eqs. (1) and (2), the cycles to failure, N, have a lognormal distribution, and the variable $\log N$ is normally distributed. A wide variety of fatigue experiments show that the lognormal distribution fits the data of cycles to failure better than other common models [2]. Note that the mean value of $\log N$ is a function of the stress level, S, but the variance of $\log N$ is a constant because the m of each curve is constant. This constant-variance assumption has been borne out in fatigue tests on a large number of steel and aluminum specimens [4]. Graphically, the *P-S-N* diagram of this model presents a family of parallel lines on log-log scales as shown in Fig. 1. It has to be pointed out that under this model both the mean and the variance of the fatigue life, N, vary with the stress levels as usual. However, the coefficient of variation of N appears to be a constant regardless of the stress levels, then

$$C_N = C_A, \tag{3}$$

where C_A is the coefficient of variation of A, and is obtained from test data. Therefore, the distribution of fatigue life, under the constant-amplitude stress with the known magnitude S, can be written as [5]

$$f(N) = \frac{1}{N\,\sigma_{N'}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log_e N - \overline{N'}}{\sigma_{N'}}\right)^2}, \tag{4}$$

where the parameters, $\overline{N'}$ and $\sigma_{N'}$, are the mean and standard deviation of $\log N$, respectively. They are obtained from the *P-S-N* model, or

$$\overline{N'} = \overline{A'} - m \log S, \tag{5}$$

and

$$\sigma_{N'} = \sigma_{A'}. \tag{6}$$

The reliability function, then, is given by [5]

$$R(N_1) = \int_{N_1}^{\infty} f(N)\,dN = 1 - \Phi\left(\frac{\log_e N_1 - \overline{N'}}{\sigma_{N'}}\right), \tag{7}$$

where

N_1 = cycles of operation.

For a lot of materials, this model provides a good approximation to the cycles to failure, or fatigue life under constant-amplitude loading. However, a variety of stress types may be encountered in the real world. Hence, the practical problem is how to use the abundance of available constant amplitude data for fatigue reliability analysis under various conditions [2].

FATIGUE LIFE DISTRIBUTION UNDER FIXED AMPLITUDE STRESS: In engineering design practice, some mechanical components may be subjected to a fully oscillatory stress with a nearly fixed amplitude during service. But there is uncertainty associated with the magnitude of the stress amplitude. In other words, the value of the stress amplitude, S, does not vary with working time, but is a random variable at the design stage. In the preceding section, the magnitude of the stress amplitude was assumed to be known exactly. Therefore, the reliability, given by Eq. (7), is in reality the conditional reliability in this case, or

$$R(N_1|S) = \int_{N_1}^{\infty} \frac{1}{N\,\sigma_{N'}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log_e N - \overline{N'}}{\sigma_{N'}}\right)^2} dN. \tag{8}$$

If the stress amplitude distribution is given by $f(S)$, then, by the Theorem of Total Probability [11, pp. 8], the reliability function of a component can be written as

$$R(N_1) = \int_S R(N_1|S)\,f(S)\,dS. \tag{9}$$

Finally, the *pdf* of the fatigue life is found by differentiating Eq. (9) with respect to N_1 leading to

$$f(N_1) = \int_S f(N_1|S)\,f(S)\,dS, \tag{10}$$

where $f(N_1|S)$ is a conditional *pdf*; i.e., the *pdf* of N_1 at a given stress level, S, which is defined by Eq. (4) as

$$f(N_1|S) = \frac{1}{N_1\,\sigma_{N'}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log_e N_1 - \overline{N'}}{\sigma_{N'}}\right)^2}. \tag{11}$$

When the stress amplitude is modeled as a normal variable, the *pdf* of fatigue life will be

$$f(N_1) = \int_{-\infty}^{+\infty} \frac{1}{N_1\,\sigma_{N'}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log_e N_1 - \overline{N'}}{\sigma_{N'}}\right)^2} \frac{1}{\sigma_S\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{S-\overline{S}}{\sigma_S}\right)^2} dS, \tag{12}$$

where $\overline{S}$ and σ_S are the mean and standard deviation of the stress amplitude, S, respectively. The typical shape of $f(N_1)$, defined in Eq. (12), is shown in Fig. 2.

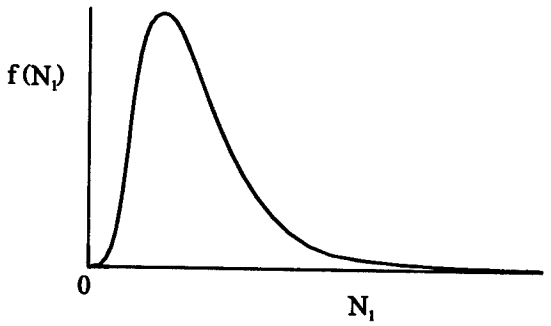


Fig. 2 – The *pdf* of the cycles to failure under a constant-amplitude stress with the associated uncertainty of the magnitude of this stress amplitude, obtained by using the following parameters: $\overline{A} = 4 \times 10^9$, $C_A = 0.50$, $\overline{S} = 40$, and $\sigma_S = 2$.

For engineering analysis convenience, the fatigue life distributions used most frequently have been the lognormal and the Weibull. Both are applicable to a wide range of materials and fatigue testing conditions. As may be seen, the shape in Fig. 2 may be well represented by either the lognormal or the Weibull distribution. The estimates of the parameters of these distributions can be obtained using the matching moments methods [5, pp. 399-416, and pp. 282-313].

FATIGUE LIFE DISTRIBUTION UNDER RANDOM STRESS: In general, the fatigue stress histories, $s(t)$, commonly observed in structural and mechanical components, are random processes. Unfortunately, the test data on random fatigue are limited. Almost all available fatigue data for design purposes are based on constant-amplitude tests. A common approach is to conduct a damage accumulation hypothesis to relate fatigue behavior under variable-amplitude stress to the "known" behavior under constant-amplitude stress. The most widely used cumulative damage rule is Miner's rule. According to the Miner's rule, damage, D, is defined as

$$D = \sum_i \frac{n(S_i)}{N(S_i)}, \tag{13}$$

where

$n(S_i)$ = number of stress cycles at stress level S_i,

and

$N(S_i)$ = number of cycles to failure at constant-amplitude stress, S_i.

Fatigue failure occurs when, on the average, $D \geq 1$.

It is generally thought that Miner's rule works reasonably well for random-load fatigue. Some studies have shown that Miner's rule yields quite an accurate estimate for the mean life of a component subjected to fatigue [10].

Mean Value of Fatigue Life Under Random Stress: To use Miner's rule, designers have to know how to count the stress cycles, the number, $n(S_i)$, and the amplitude, S_i, of the stress cycles. It is easy when the stress history is a stationary narrow-band process with zero-mean. The narrow-band process is characterized by its approximate constant period, or frequency. The midpoint of each single cycle of stress is equal to the mean of stress. Hence, the stress peak can be used equally as stress amplitude and the rate of zero-up-crossing as the frequency of stress cycles. If the fatigue stress is a stationary narrow-band Gaussian process with zero-mean, the distribution of the stress peaks, hence of the stress amplitude, is the Rayleigh distribution [3], and the mean of the fatigue life is given by

$$\overline{T} = \frac{\overline{A}}{n^+(0)\,(\sqrt{2}\sigma_s)^m\,\Gamma\left(\frac{1}{m}+1\right)}, \tag{14}$$

where

$n^+(0)$ = expected zero-up crossing rate of the stress process [12, p. 154],

or

$$n^+(0) = \sqrt{\frac{\int_0^\infty f^2 W(f)\,df}{\int_0^\infty W(f)\,df}},$$

$\overline{A}$ = mean value of A,

σ_s = rms of the stress process = $\int_0^\infty W(f)\,df$,

and

$W(f)$ = spectral density function of the stress process.

When the stress process is wide-band, stress-cycle counting becomes difficult. The value of the stress peak can not be used as stress amplitude even though the mean of the stress process is zero. It is not obvious exactly what a cycle is and how it should be counted to be used with Miner's rule. Among fatigue experts, the rainflow method is generally regarded as the method leading to a better prediction of fatigue life. The rainflow method determines the number and the amplitude of all stress cycles from a finite sample of recorded stress history of hot spots in the component. However, it becomes impossible in the case when stress records can not be obtained, for example at the design stage.

For a wide-band Gaussian stress, under some assumptions, the joint distribution of the amplitude and the mean of a single stress cycle is derived, based on the "half-range counting method", as follows:

Consider a part of the stress sample shown in Fig. 3. A single cycle is started at a valley such as Valley 1. Each trough-originated stress range (i.e., a path from a valley to the next adjacent peak) is considered as a half-cycle. Similar to the rainflow method, it is assumed that, for a stress history sufficiently long, any "trough-originated" half-cycle can be paired with a "peak-originated" half-cycle of the same range somewhere to form a single stress cycle. As shown in Fig. 3, a half cycle from Valley 1 to Peak 2 is paired with a half cycle from Peak 6 to Valley 7 to form a single stress cycle with the range of 40 and the mean value of 60. Therefore, the range and the mean value of a stress cycle should be counted

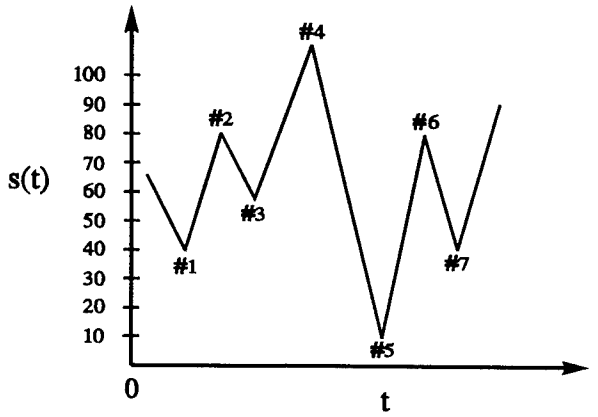


Fig. 3 – A part of the time history of a wide-band stress.

on the trough-originated stress path only, and the paired peak-originated half cycle always appears later in the stress process; i.e., the rise, defined as the difference between the stress at a valley and the stress at the next peak, is considered as the range of a single stress cycle, and the midpoint between this adjoining valley and the peak is considered as the mean value of this single stress cycle.

If the stress is a stationary, wide-band, Gaussian process with zero-mean, the joint density function of a trough-originated range, S_r, and its mean value, S_m, is given by [12, p. 170]

$$f_{rm}(S_r,\ S_m) = \frac{S_r}{4\,\alpha^2\,\sigma_s^2} \exp\left(-\frac{S_r^2}{8\,\alpha^2\,\sigma_s^2}\right) \frac{1}{\sqrt{2\,\pi}\,\epsilon\,\sigma_s} \exp\left(-\frac{S_m^2}{2\,\epsilon^2\,\sigma_s^2}\right), \tag{15}$$

where

$$\alpha = \text{irregularity factor of the stress process} = \frac{\int_0^\infty f^2\,W(f)\,df}{\sqrt{\int_0^\infty W(f)\,df\ \int_0^\infty f^4\,W(f)\,df}},$$

and

$$\epsilon = \text{spectral width parameter of the stress process [1]} = \sqrt{1-\alpha^2}.$$

Note, from Eq. (15), that the range and the mean value of the stress cycle are independent, because their joint *pdf* is factorable, and the *pdf*'s of S_r and S_m can be obtained by

$$f_r(S_r) = \int_{-\infty}^{\infty} f_{rm}(S_r,\ S_m)\,dS_m = \frac{S_r}{4\,\alpha^2\,\sigma_s^2} \exp\left(-\frac{S_r^2}{8\,\alpha^2\,\sigma_s^2}\right), \tag{16}$$

and

$$f_m(S_m) = \int_0^{\infty} f_{rm}(S_r,\ S_m)\,dS_r = \frac{1}{\sqrt{2\,\pi}\,\epsilon\,\sigma_s} \exp\left(-\frac{S_m^2}{2\,\epsilon^2\,\sigma_s^2}\right), \tag{17}$$

respectively.

Since the stress amplitude, S, is half of the range of the stress cycle, S_r, the *pdf* of the stress amplitude is given by

$$f_a(S) = |\frac{dS}{dS_r}| f_r(2S) = \frac{S}{(\alpha \sigma_s)^2} \exp\left[-\frac{S^2}{2(\alpha \sigma_s)^2}\right]. \tag{18}$$

Note that, from Eqs. (17) and (18), the stress cycle amplitude is a Rayleigh variable and the stress-cycle mean is normally distributed with a zero mean. So, the stress cycle mean will, usually, not be zero even if the mean of the whole stress process is zero. As usual, the stress-cycle mean is handled using the modified Goodman diagram, or

$$N = A\left(\frac{S}{1 - S_m/S_U}\right)^{-m}, \tag{19}$$

where

S_U = ultimate strength.

If the same derivation procedure, as in the narrow-band situation, is employed, then, the mean value of the fatigue life is given by

$$\overline{T} = \frac{\overline{A}}{n_p^+ (\sqrt{2}\alpha \sigma_s)^m \Gamma\left(1 + \frac{1}{m}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\epsilon\sigma_s} \frac{1}{(1 - S_m/S_U)^m} \exp\left[-\frac{S_m^2}{2(\epsilon\sigma_s)^2}\right] dS_m}, \tag{20}$$

where

$$n_p^+ = \text{expected peak frequency} = \sqrt{\frac{\int_0^\infty f^4 W(f)\, df}{\int_0^\infty f^2 W(f)\, df}}.$$

In general, the hot-spot-stress process will have a nonzero mean in mechanical and structural systems. Hence, the stress-cycle mean has the *pdf*

$$f_m(S_m) = \frac{1}{\sqrt{2\pi}\epsilon\sigma_s} \exp\left[-\frac{(S_m - \mu_s)^2}{2(\epsilon\sigma_s)^2}\right], \tag{21}$$

where

μ_s = mean of the whole stress process.

Finally, the mean value of the fatigue life under a stationary Gaussian stress process is given by

$$\overline{T} = \frac{\overline{A}}{n_p^+ (\sqrt{2}\alpha \sigma_s)^m \Gamma\left(1 + \frac{1}{m}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\epsilon\sigma_s} \frac{1}{(1 - S_m/S_U)^m} \exp\left[-\frac{(S_m - \mu_s)^2}{2(\epsilon\sigma_s)^2}\right] dS_m}. \tag{22}$$

This model requires the stress spectral density function, $W(f)$, only. A stress history is not required. Note that the static property of the material, S_U, is treated as a deterministic variable here. But, it can also be treated as a random variable when the distribution of S_U is available.

Standard Deviation of Fatigue Life Under Random Stress: Since considerable statistical scatter exists in the cycles to failure data, determination of the variance of the fatigue life should attract a great deal of attention. Some approximations, based on Miner's rule and stochastic process theory, have been published for a stationary, narrow-band Gaussian

stress [1, 9]. An alternate solution is to calculate the coefficient of variation of the fatigue life, C_T.

In analyzing a large number of variable-amplitude fatigue test data on steel and aluminum, it has been found that the scatter under random loading, C_{N_f}, is nearly identical with the scatter in constant-amplitude tests, C_A, [6, 7, 8]; i.e.,

$$C_T \cong C_A. \tag{23}$$

Hence, the standard deviation of the fatigue life can by estimated by

$$\hat{\sigma}_T = C_A \overline{T}, \tag{24}$$

where $\overline{T}$ is given by Eq. (22).

Fitigue Life Distribution Under Stress: In general, the random fatigue problem is quite complicated, and fatigue reliability analysis must rely on assumed distributions. As said before, the distributions for fatigue life under random loading, widely used in engineering, are the lognormal and the Weibull. For example, under the lognormal assumption, the estimated *pdf* of the fatigue life will be

$$f_{LN}(T) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_{T'} T} \exp\left[-\frac{1}{2}\left(\frac{\log_e T - \widehat{\overline{T'}}}{\hat{\sigma}_{T'}}\right)^2\right]. \tag{25}$$

Then, the reliability function is

$$R_{LN}(T) = \int_T^\infty f_{LN}(x)\,dx = \Phi\left(\frac{\widehat{\overline{T'}} - \log_e T}{\hat{\sigma}_{T'}}\right), \tag{26}$$

and the failure rate function is

$$\lambda_{LN}(T) = \frac{f_{LN}(T)}{R_{LN}(T)}. \tag{27}$$

The estimates of the parameters may be obtained using the matching moments method, or from

$$\widehat{\overline{T'}} = \log_e \frac{\overline{T}}{(1 + C_T)^{1/2}},$$

and

$$\hat{\sigma}_{T'} = [\log_e(1 + C_T^2)]^{1/2}.$$

Once the fatigue reliability functions are available, designers can determine the safe operating life of a component, schedule necessary preventive maintenance and repairs, select an allowable load spectrum, develop design criteria, and so on.

Numerical Example: The fatigue stress in the hot-spot of a mechanical component is found to be a stationary wide-band Gaussian process with the following parameters:

$\mu_s = 9.0$ ksi, $\sigma_s = 5.0$ ksi, $\alpha = 0.6$, and $n_p^+ = 15$/sec.

Table 1 – Summary of the results for the example.

No.	Results	
1	Mean and standard deviation of the fatigue life	
	$\overline{T} = 5,364.67$ hr	$\hat{\sigma}_T = 1,341.17$ hr
2	Lognormal parameter estimates	
	$\widehat{\overline{T'}} = 8.557$	$\hat{\sigma}_{T'} = 0.246$
	Weibull parameter estimates	
	$\hat{\beta} = 4.47$	$\hat{\eta} = 5,880.98$ hr
3	Mission reliability	
	Lognormal: $R_{LN}(3,000 \text{ hr}) = 0.9874$	
	Weibull: $R_W(3,000 \text{ hr}) = 0.9518$	

The *P-S-N* model of the material used is given by

$$m = 3,272, \quad \overline{A} = 5.012 \times 10^{10}, \quad \text{and} \quad C_A = 0.25.$$

Do the following:

1. Determine the mean and standard deviation of the fatigue life of this component.

2. Estimate the parameters of its fatigue life distributions under the lognormal and Weibull assumptions, respectively.

3. Evaluate the reliability of this component in a mission of 3,000 hr under the lognormal and Weibull assumptions, respectively.

Solutions To Example: Substituting the given information into Eqs. (22) and (24), applying the matching moments method, and using the lognormal and Weibull reliability functions [5] yield the results in Table 1. The obtained *pdf*'s of the fatigue life are shown in Fig. 4.

References

[1] Sun, Feng-Bin and Kececioglu, Dimitri B., "Fatigue Reliability Evaluation and Preventive Maintenance Policy for a Structural Element Under Random Vibration," presented at the 37th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, April 15-17, 1996, Salt Lake City, Utah, and published in its Proceedings, pp. 2167-2174, 1996.

[2] Sundararajan, C., *Probabilistic Structural Mechanics Handbook*, Prentice Hall, Englewood Cliffs, NJ, 740 pp., 1996.

[3] Wirsching, P. H., *Probabilistic Mechanical Design*, Lecture Notes, The University of Arizona, 328 pp., 1995.

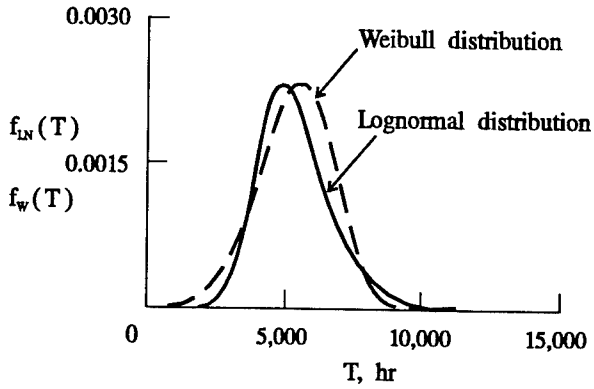


Fig. 4 – The obtained *pdf* of fatigue life under the lognormal and the Weibull distributions assumptions.

[4] American Society of Civil Engineers (ASCE), Committee on Fatigue and Fracture Reliability (1982), "Fatigue and Fracture Reliability: A State-of-the-art Review," *Journal of the Structural Division of American Society of Civil Engineers,* Vol. 108, pp. 3-104, 1982.

[5] Kececioglu, Dimitri B., *Reliability Engineering Handbook,* Prentice Hall PTR, Upper Saddle River, NJ, Vol. 1, 720 pp., Fifth Printing in 1996.

[6] Jacoby, G. H., and Nowack, H., "Comparison of Scatter Under Program and Random Loading and Influencing Factors," *Probabilistic Aspects of Fatigue, ASTM STP 511,* American Society for Testing and Materials, pp. 61-74, 1972.

[7] Whittaker, I. C., and Besuner, P. M., "A Reliability Analysis Approach to Fatigue Life Variability of Aircraft Structures," *Wright-Patterson Air Force Base Technical Report, AFML-TR-69-65,* Wright-Patterson AFB, Ohio, April, 1969.

[8] Wang, Wendai "Digital Simulation of Random Load Spectra with Respect to Fatigue," *Thesis Presented to Shanghai Jiaotong University in Partial Fulfillment of the Requirements for the Degree of Master of Science,* 62 pp., Shanghai, China, 1987.

[9] Zhu, W. Q., and Lei, Y., "A Stochastic Theory of Cumulative Fatigue Damage," *Probabilistic Engineering Mechanics,* Vol. 6, pp. 222-227, 1991.

[10] Saunders, S. C., "On the Probabilistic Determination of Scatter Factors Using Miner's Rule in Fatigue Life Studies," *Probabilistic Aspects of Fatigue, ASTM STP 511,* American Society for Testing and Materials, pp. 185-203, 1972.

[11] Spiegel, Murray R., *Probability and Statistics,* McGraw-Hill Book Co., 1221 Avenue of Americas, New York, NY 10020, 372 pp., 1975.

[12] Wirsching, P. H., Paez, T. L., and Ortiz, K., *Random Vibrations: Theory and Practice,* John Wiley & Sons, 488 pp., 1995.

NODE-BASED STRUCTURAL SHAPE OPTIMIZATION

X. Song and J.D. Baldwin
School of Aerospace and Mechanical Engineering
University of Oklahoma
865 Asp Avenue, Room 212
Norman, OK 73019

Abstract: A novel, node-based shape optimization method is developed for planar structures based on the consideration that the critical stress and displacement constraints are generally located along or near the structural boundary. The proposed method puts the maximum weights on the selected boundary nodes, referred to as the design points, so that the time-consuming sensitivity analysis is based on the perturbation of only these nodes. The method also allows large shape changes to achieve the optimal shape. The design variables are specified as the moving magnitudes for the prescribed design points which are always located at the structural boundary. The paper emphasizes triangular finite element solution technique which is consistent with automatic mesh generation; an adaptive mesh refinement process is presented and implemented for triangular meshes. Furthermore, structural weight, which in most applications is the objective function, is perturbed only if the structural boundary changes. The versatility of the new method is demonstrated in terms of examples.

Key Words: Adaptive mesh refinement; Design sensitivity analysis; Error estimation; Finite element analysis; Shape optimization

INTRODUCTION: The efficient use of materials is a primary concern in engineering design. To this end, the subject of combining the finite element (FE) structural analysis technique and numerical optimization algorithms has been widely investigated in the last two decades [1], and subsequently has been incorporated into a number of commercial finite element packages. Among many structural optimization methods, only numerical shape optimization, or optimal shape design, will be discussed in this work. Simply speaking, the shape optimization design uses numerical optimization techniques in an attempt to achieve the "best" shape (and thus the most efficient use of materials) for a structure under various constraints imposed by the design conditions.

It has been reported [2–3] that two key elements in developing a successful shape optimization algorithm are to represent efficiently the structural boundary and to predict accurately its optimal movement. Imam [4] proposed the design element concept to divide a structure into a number of substructures (design elements) which are allowed to be altered to yield the optimal shape. One benefit offered by this methodology is that the initially coarse mesh can be refined locally in each design element. Its disadvantages are obvious too. For instance, it may be difficult to establish a fully automatic process because the number and the locations of the predetermined design elements may be required to change to cope with the varying boundary shapes. Furthermore, incorporation of a local refinement within a single design element into the global model generally imposes a major bookkeeping load. Consequently, application of the shape optimization procedure based on the design element concept must in general be restricted to designs anticipating only small shape changes. Accurate prediction of the displacements and the stresses is also required for a successful shape optimization. An efficient way of achieving this is to use adaptive mesh refinement (AMR) in which error estimation and automatic mesh generation are combined so that the mesh is refined only in the regions where error estimation indicates relatively large errors from the FE solution. The difficulty of

including AMR technique in the shape optimization was mentioned by Belegundu and Rajan [5]. To circumvent this complexity, they develop a method based on fictitious loads acting on an auxiliary structure. One of the purposes in employing this auxiliary system is to provide smooth mesh translations for the varying structural shapes during optimization iterations by altering the preceding meshes, while seeking to avoid an overall structural remesh. It should be pointed out that the limited experience from these authors suggests that a periodic remesh is necessary to avoid excessive distortion in the mesh system.

This paper reports the development of a node-based shape optimization methodology. The boundary nodes and elements are used to prescribe design variables on which the derivatives of the objective and constraint functions are evaluated to invoke the mathematical programming method of optimization. A modified version of Vanderplaats' constrained minimization routine CONMIN [6] is used to find optimal structural configurations. In this way, the design variables can be explicitly specified in a natural way on the boundary.

THEORETICAL DEVELOPMENTS: It was observed that moving each design point individually as dictated by the optimization solver can lead to unrealistic shapes which may contain kinks. To remedy this, a smoothing process, the least square B-spline fitting method, is used. The AMR technique is invoked to provide more accurate stress field. A compact sensitivity expression of stress constraint equation is also presented. These topics are briefly summarized as follows.

Least Square B-Spline Fitting: It will be demonstrated in the next section that a small number of boundary nodes in a FE model can be selected as the design points. In this way, the number of design variables is reduced dramatically while retaining flexibility of allowing the boundaries to move. After each optimization iteration, a boundary translation suggestion is made for each design point. Due to the fact that the number of boundary nodes in the FE model is generally far greater than the number of design points, it is necessary to predict the movement for nodes other than design points. Interpolation or extrapolation may be used for this purpose. It has been found, however, that a robust procedure can be constructed using the B-spline fitting technique.

A least square B-spline fitting procedure consists of the following two steps. First, the control polygon points are computed by using a least square solver; the resulting polygon points can then be used to approximate the desired curve. The procedure can be developed from techniques outlined by Rogers and Adams [7]. In this study, only the uniform open knot vector is used to evaluate B-spline basis functions.

Error Estimation and AMR: In general, the calculated element stresses and strains based on the FE solution are discontinuous at the interface of two elements. For a homogeneous domain, however, it is expected that the exact stress and strain should be continuous functions across elements. The calculation of more accurate stress solutions is one of the most active topics in FEM in recent years. A promising method, superconvergence patch recovery (SPR), was proposed in [8] and is adopted in this study. Simply speaking, SPR leads to a superconvergent estimate for the smoothed stress at an apex node by patching together the elements surrounding the node and employing a local least-squared fitting technique.

A patch employed here is generally represented by a second-order complete polynomial $\{1,x,y,x^2,xy,y^2\}^T$. If the number of integration points associated with the patch is not less than six or if the interior integration scheme is used. Otherwise, the following reduced complete polynomial is used $\{1,x,y\}^T$. It is worth mentioning that no apparent advantage was obtained by using the third or higher-order complete polynomials. To avoid the

potential ill-conditioning in using the least square fitting calculations, the integration points associated with a patch are linearly translated into a two by two square centered at point (0,0). Also, stresses are normalized to maintain the same magnitude on both sides of the least square equation.

Only *h*-refinement is considered here for mesh refinement. A two-dimensional Delaunay triangulator [9] is used to fulfill the triangular mesh generation as well as *h*-refinement.

Design Sensitivity Analysis: One of the main objectives in structural design is to prevent failure, that is, the stresses must be within safe limits. Two strength criteria used against the limits are based on either Tresca's maximum shear stress theory or the Von Mises effective stress theory. In matrix notation, these stresses, for a planar problem, can be represented by

$$\sigma_\chi^2 = \{\sigma\}^T [V_\chi] \{\sigma\} \tag{1}$$

where the subscript χ can take *SH* or *VM* to represent the maximum shear stress or the Von Mises stress, respectively. The corresponding matrices are expressed as

$$[V_{\mathrm{SH}}] = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad [V_{\mathrm{VM}}] = \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \tag{2}$$

In this work, the strength constraints are expressed by either one of

$$g = \sigma_\chi^2 / \sigma_A^2 - 1 \tag{3}$$

where σ_A^2 is the square of the prescribed failure value, which in general can be either one of σ_C^2, σ_T^2 and $\sigma_T \sigma_C$ (σ_T and σ_C denote the allowable stresses for tensile and compressive failure, respectively.) It can be shown for a pointwise strength constraint that

$$\frac{dg}{dp} = \frac{2}{\sigma_A^2} \{\sigma\}^T [V_\chi][D] \left([B] \frac{\partial \{u\}}{\partial p} + \frac{\partial [B]}{\partial p} \{u\} \right) \tag{4}$$

The sensitivity results of strength constraints in the integral form are also used in this study. The strength constraint in the area integral form is defined for each element by

$$\bar{g} = \frac{1}{A_e} \int_{A_e} g \, dA = \frac{1}{A_e} \int_{\eta_1}^{\eta_2} \int_{\xi_1}^{\xi_2} \sigma_\chi^2 / \sigma_A^2 |[J]| \, d\xi d\eta - 1 \tag{5}$$

The sensitivity with respect to a design variable p can be found as

$$\frac{d\bar{g}}{dp} = \frac{1}{A_e} \left[\int_{A_e} \frac{dg}{dp} \, dA - \bar{g} \int_{\eta_1}^{\eta_2} \int_{\xi_1}^{\xi_2} \frac{d|[J]|}{dp} \, d\xi d\eta \right] \tag{6}$$

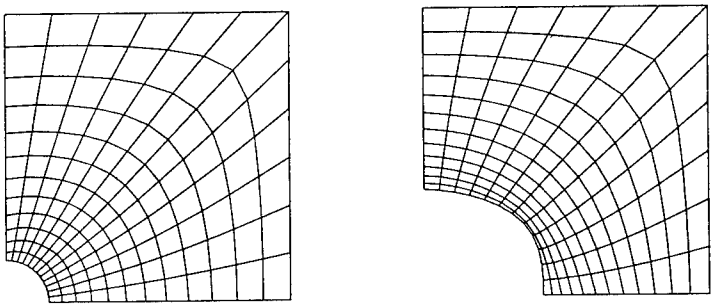


Figure 1: Illustration of varying meshes whose topology is fixed; initial mesh (left) and mesh after 15 iterations (right)

PROPOSED SHAPE OPTIMIZATION METHOD: Proper specification of design variables is very important for successful shape optimization and there are many ways to do it. The most popular methods include the design element approach [4] in which the underlying structure is divided into a number of substructures on which the overall finite element model is based. The design variables are first specified at the substructure level, and then other necessary information comes from interpolating by shape functions, for example. This approach is generally restricted to problems which are expected to achieve the optimal shape with a small shape change since a large shape change may result in severe distortion of the boundary elements and lead to unreliable FE solutions.

As mentioned earlier, difficulties arise when mesh topology varies during the optimization process due to adaptive modeling, as it inevitably alters the number of nodes and elements and their ordering. It can be argued, however, that the adaptive modeling is inevitable for most of the structural shape optimization applications for at least two reasons. First, during the iterations, the boundaries may vary to such an extent that problems associated with excessive element distortions and unreasonably large aspect ratios can occur in some elements if the mesh topology is kept fixed. Figure 1 shows such an example, encountered by in the course of this study. The meshes (initial and after 15 iterations) consist of eight-node quadrilateral elements. As is clearly shown, a number of elements in the later mesh have large aspect ratios which would be expected to lead to inaccurate prediction of the stresses. Second, it is generally required for an adaptive FE model of a large structure to accurately and economically predict the locations and values of critical stress. A uniformly fine mesh is very costly and unnecessary. A coarse mesh will underestimate the stress value in a average sense, especially near the boundaries. Thus, adaptive meshing is also valuable in this respect.

In this work, the design variables are associated with a limited number of selected boundary nodes. The specification procedure of these design variables is presented and its implementation is discussed. The final section contains a flow chart which outlines the main steps necessary to accomplish an integrated shape optimization analysis. It should be pointed out that the node-based shape optimization concept is not new. In fact it was proposed by Zienkiewicz and Campbell [10]. Since then many applications have appeared in the literature and many improvements and new methods have been proposed and applied. As far as we know, all applications of the node-based shape optimization technique require important information, such as sensitivity analysis results, at all nodal points. This is obviously unnecessary due to the fact that the critical constraints are almost always located along or near the structural boundary. Furthermore, structural

weight, which in most applications is the objective function, is perturbed only if the structural boundary changes. Based on this consideration and using a technique from the boundary layer concept in viscous fluid dynamics, the proposed method puts the maximum weights on the boundary nodes, referred to as the design points, so that the time-consuming sensitivity analysis is based only on the perturbation of these nodes.

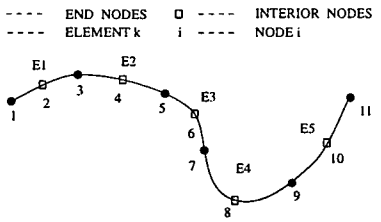


Figure 2: Illustration of a segment of quadratically one-dimensional elements

Design Variables Specification: It is proposed that only boundary information (nodal point coordinates and types of elements) is used to specify all the design variables necessary for a shape optimization design. For a boundary node, its coordinates, $\vec{r} = (x, y)^T$, define a single design variable p as

$$\vec{r} = \vec{r}_0 + p\vec{s} \tag{7}$$

Hereafter, the vector $\vec{s}$ is referred to as the nodal moving velocity which can be either predetermined or recalculated at each step. Most of the optimization studies reported in the literature employed the predetermined moving velocity methodology. The major disadvantage of this method is that it discourages the convergence of the iterative process because, as the boundary varies, the predetermined velocity may have its direction parallel to the boundary segment to be changed to achieve an optimal shape.

The moving velocities used in this study are recalculated at every iterative step. It is of interest to note that the nodal moving velocity can be simply expressed by

$$\vec{s} = \frac{\partial \vec{r}}{\partial p} \tag{8}$$

Explicit expressions can be developed for moving velocities at boundary nodes as follows. Suppose that a segment of the structural boundary consists of a number of quadratic curve elements, as depicted by Figure 2. A boundary node in this segment is identified to be either the interior node or an end node of a quadratic element. It is noted that a quadratic element consisting of three nodes *i, j,* and *k* in an adequate sequence can be approximated by

$$\begin{aligned} x(\xi) &= N_i(\xi)x_i + N_j(\xi)x_j + N_k(\xi)x_k \\ y(\xi) &= N_i(\xi)y_i + N_j(\xi)y_j + N_k(\xi)y_k \end{aligned} \tag{9}$$

where $N_i(\xi)$ and (x_i, y_i) are the shape function and the coordinates of the i-th node. Similar expressions can be developed for other two nodes. As a result, the normal vector at an arbitrary point of this element can be represented by

$$\vec{n} = \left(y'(\xi), -x'(\xi)\right)^T \tag{10}$$

This vector is readily used to determine the nodal moving velocities. For a node which connects to only a single element, the velocity $\vec{s}$ is simply identified to be the normal vector. For end nodes which connect to two elements, the average of two normal vectors obtained from two connected elements is employed to yield the velocity. A nodal moving velocity generally includes both magnitude and direction. The magnitude may be absorbed into the design variable p so that $\vec{s}$ can be always assumed as a normalized vector.

We have seen that a design point is associated with a single design variable p. The number of design variables can be further reduced by decreasing the number of design points. It have been a common belief that curved elements are not recommended for FE solutions. Even if it is absolutely necessary to use them, interior nodes should be the midpoints of end nodes, except for modeling a structure with a singularity. Consequently, it is further proposed that only end nodes are selected as the candidates for design points. As a result of this simplification, each design variable perturbs geometry of only one or two boundary elements, although its effects may likely be propagated into other elements by subsequent automatic mesh generation.

The sensitivities of the area and components of the force vector and the stiffness matrix with respect to p can be expressed as a summation of the product of the nodal moving velocities and the derivatives of the functions with respect to the nodal point coordinates by

$$\frac{\partial f}{\partial p} = \sum \left(\frac{\partial f}{\partial \vec{r}_i} \right)^T \frac{\partial \vec{r}_i}{\partial p} = \sum \left(\frac{\partial f}{\partial \vec{r}_i} \right)^T \vec{s} \tag{11}$$

It is noted that the vector $\partial f / \partial \vec{r}_i$ is zero if the coordinates $\vec{r}_i$ are not associated with design points. Consequently, these are only a few nonzero sensitivities of components of the force vector and the stiffness matrix. On the other hand, the sensitivities of constraints involving the displacement are likely nonzero through the solution $\partial \{u\} / \partial p$.

Implementation: To implement the present method, it is necessary to select a number of nodes as design points. As stated previously, these design points should always be on the structural boundary. The structural boundary is first divided into a certain number of boundary segments. Each segment may be regarded as a design segment which is susceptible to modification in finding an optimal structural shape. Once design segments are identified, a certain number of nodal points are assigned to each segment as design points. The number of design points on each segment is initially prescribed and fixed in subsequent iterations. It should be pointed out that other boundary segments also need to be supplied with nodal points for meshing purposes. These boundary nodes can then be used to establish an initial mesh system by the automatic mesh generator.

After each iteration, the modified boundary segment is smoothed by the least square B-spline fitting technique. This smoothing technique requires inputs of data points and the number of control polygon points (order of used B-spline basis functions is not regarded as another parameter since K=4 is used exclusively in this study). In this study, only design points are used to provide the required data points. Also, the number of control polygon points is provided at the beginning of the iterations. A suitable number depends on how the smoothed curve behaves. An excessive number of control polygon points will return a curve with excessive number of kinks. On the other hand, an underestimated number results in a curve far from the predicted optimal shape. It was found that in general a number between one-third and a half of the number of data points is adequate.

RESULTS AND DISCUSSIONS: Two examples of the shape optimization algorithm are presented here.

Cantilever Beam: The first example problem we consider is to determine the optimal shape of a tip-loaded cantilever beam. This problem serves as an optimal shape verification example given that its analytic solution is available. The configuration and the FE model associated with the cantilever beam are shown in Figure 3, in which the distributed load is specified as p=34.475 Pa (5 psi) and the uniform thickness t=12.7 cm (0.5 in.). The beam is composed of an isotropic material with an elastic modulus E=20.685 GPa (3×10^7 psi) and a Poisson ratio $\nu = 0$; the maximum allowable Von Mises stress is $\bar{\sigma}_{VM} = 4.62$ MPa (670 psi), which is slightly greater than the maximum normal stress 4.55 MPa (660 psi) in the initial beam predicted by the classic beam theory. The Poisson ratio is set to zero to remain consistent with the classical beam theory which neglects the Poisson effect.

Twelve design points are assigned along the bottom boundary of the beam. A constraint that allows no movement in the x-direction is imposed on the design point at the tip. Also, y-coordinates of all design points are restricted so that they are not greater than 0.950, i.e., the beam thickness cannot vanish. Six polygon points are used for smoothing by cubic B-splines for the current problem.

It is known that σ_x, the normal x-stress, at the top and bottom edges is the main concern as to whether the Von Mises stress criteria is reached. As a result the Von Mises stress near the top and bottom boundaries can be well approximated by the absolute value of σ_x, that is,

$$\sigma_{VM} = \left| \frac{M(x)y}{2I} \right| \tag{12}$$

where $M(x)$ and I represent the bending moment and the moment of inertia, respectively. The optimal shape for the bottom boundary can be approximated from Equation (12) by

$$y = 1 - \sqrt{\frac{6|M(x)|}{t\,\bar{\sigma}_{VM}}} \tag{13}$$

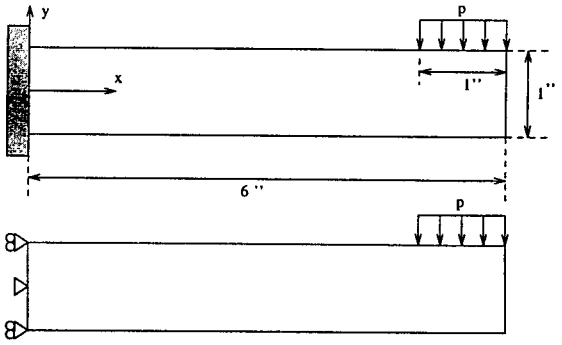


Figure 3: Plane stress model of a cantilever beam showing idealized boundary conditions

The optimum shapes of the bottom boundary from the theoretical solution, Equation (13), and the numerical result are compared in Figure 4. Only few percent difference in profile

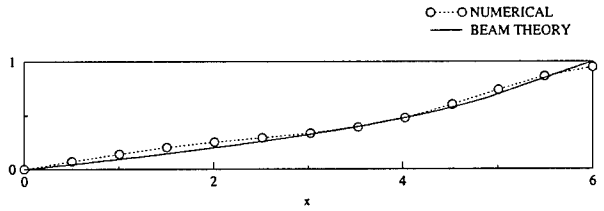


Figure 4: Comparison of optimal cross sections

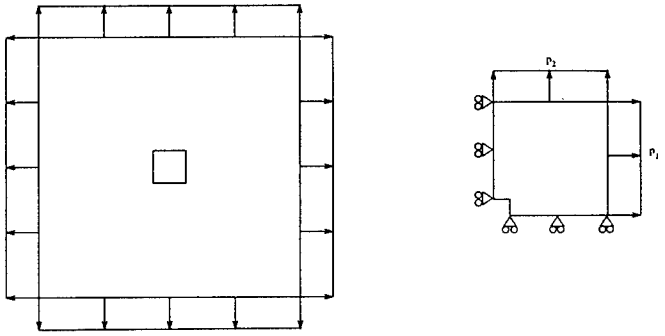


Figure 5: Plane stress model of plate with a square hole

dimension is observed. Also note that theoretical solution vanishes at the end. Compared to the popular design element method (cf. [2]), it can be seen that relatively little effort is required to set up a structural optimization process. Once the computational framework is in place, the analyst needs only specify the initial geometry and design points. In general, it is far easier to locate the design points than to set up design element models and to prescribe moving directions for an arbitrarily shaped structure.

Flat Plate with Hole: This classical problem has been widely investigated theoretically and numerically in the literature (e.g., [11], [12]). A thin 203.2 by 203.2 cm (80 by 80 in.) square flat plate with a central square hole of 30.48 by 30.48 cm (12 by 12 in.) and its finite element quarter model are shown in Figure 5. The plate is made from an isotropic material with E= 68.95 GPa (10^7 psi), ν=0.3, and $\bar{\sigma}_{VM} = 241.325$ MPa $(3.5 \times 10^4$ psi). These parameters are set identical to those used by Naqib et al. [13]. The plate is assumed to have unit thickness. A biaxial stress field is applied by specifying p_1=103.425 MPa (15,000 psi) and p_2=68.95 MPa (10,000 psi) in the x- and y-directions, respectively.

The analytical solution for the optimum shape of the hole with an infinite plate under a biaxial loading condition was summarized by Kristensen and Madsen [11], where it was shown that the hole should be an ellipse with the axis ratio *b/a* equal to the ratio of biaxial stresses, that is, $b/a = p_2/p_1$, where a and b represent the axes of the ellipse. Note, however, that this conclusion was made for the infinite plate and by ignoring possible edge effects which occur when the hole axes are comparable to the plate length. Chong and Pinter [14] addressed edge effects for a circular hole under the uniaxial loading conditions using FE solutions. They found that in general for the x-direction tensile load, as the hole diameter increases, the maximum σ_x by which the stress concentration is determined increases too, while the maximum $|\sigma_y|$ decreases.

Seventeen design points are assigned along the hole. Symmetry constraints that allow no movements in the x- and y-directions are imposed on the design points on the y and x symmetric lines, respectively. Seven control polygon points are used in the cubic B-spline smoothing.

The numerical optimization starts with 15.24 by 15.24 cm (6 by 6 in.) square quarter hole, as shown in Figure 5, where a high stress concentration is expected. During the first few iterations, the optimization routine tried to smooth the hole as much as possible. After 5 steps, the smoothed hole shape and FE mesh can be observed in Figure 6(a). The changes in the hole shape during the next 90 iterations are shown in Figures 6(b)-(d). The major and minor axes ratio for the final shape is was 1.43, which is only a 4.7 percent error compared with the analytical result. A contour plot of the Von Mises stress at the final step shows that the lack of a clear local stress concentration is confirmation that material has been allocated most efficiently to carry the applied loading.

It should be mentioned that in two other runs with the initial holes being either a circular of 15.24 cm (6 in.) radius or a very irregular shape, little difference was observed between the final computed the optimal shapes. Thus, we have demonstrated that this shape optimization algorithm is very robust in dealing with large shape changes. Adaptive mesh refinement and local error estimation make this possible.

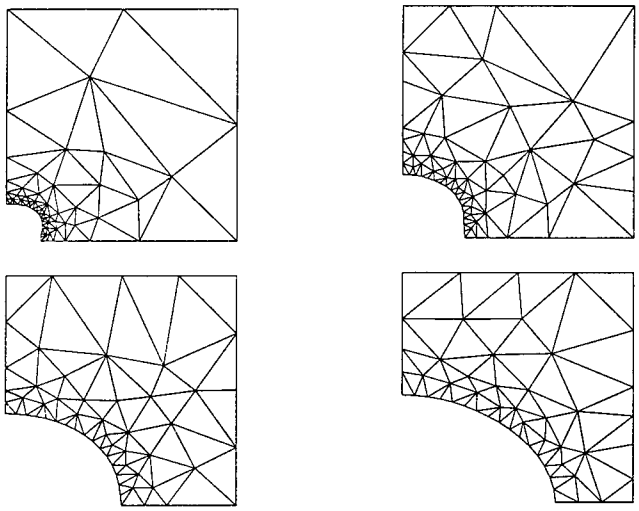


Figure 6: Snapshots and their meshes during shape optimization after iteration steps; (a) 5, (b) 35, (c) 65 and (d) 95.

CONCLUSIONS: This work presents an investigation of a novel method of specifying design variables for general structures. The method emphasizes the use of information provided by boundaries of a structure, on which the area change depends and where the critical constraints normally occur. It combines the features of flexibility of nodal points and regularity of B-spline blending functions.

The varying structural shapes are modeled via adaptive mesh refinement processes so

that the constraints can be calculated accurately and economically. This work also emphasizes the coupling among the mathematical programming solver and adaptive modeling. However, the changes of mesh topology due to the AMR cause most of the old information about the constraints and their sensitivities to be disregarded after each iteration. Although much research is needed in this aspect, the results presented in this study are encouraging.

The method developed is applied to the optimal shape solutions of several examples of plane stress structures. It can be seen that the present methodology avoids much time-consuming preprocess of specifying the proper design variables required in most existing shape optimization methods. By using the proposed method, a designer is required to provide only the boundary nodes which define overall structure and the design segments connected by a number of the boundary nodes. The input required is therefore minimum.

REFERENCES

[1] Haftka, R.T. and Gurdal, Z., *Elements of Structural Optimization*, 3rd Ed., Kluwer Academic Publisher, 1992.

[2] Ding, Y., "Shape Optimization of Structures – A Literature Survey," *Comput. Struct.*, Vol. 24(6), 985–1004 (1986).

[3] Haftka, R.T. and Grandhi, R.V., "Structural Shape Optimization – A Survey," *Comp. Meth. Appl. Mech. Eng.*, Vol. 57, 91–106 (1986).

[4] Imam, M.H., "Three-Dimensional Shape Optimization," *Int. J. Num. Meth. Eng.*, Vol. 18, 661–673 (1982).

[5] Belegundu, A.D. and Rajan, S.D., "A Shape Optimization Approach Based on Natural Design Variables and Shape Functions," *Comput. Meth. Appl. Mech. Eng.*, Vol. 66, 87–106 (1988).

[6] Vanderplaats, G.N., *CONMIN-A FORTRAN Program for Constrained Function Minimization*, NASA TM X-62282 (1973).

[7] Rogers, D.F. and Adams, J.A., *Mathematical Elements for Computer Graphics*, 3rd Ed., Prentice-Hill, Englewood Cliffs, NJ. 1985.

[8] Zhu, J.Z. and Zienkiewicz, O.C., "Superconvergence Recovery Techniques and A Posteriori Error Estimators," *Int. J. Num. Meth. Eng.*, Vol. 30, 1321–1339 (1990).

[9] Shewchuk, J.R., *A Two-Dimensional Quality Mesh Generator and Delaunay Triangulator*, Version 1.1, 1995.

[10] Zienkiewicz, O.C. and Campbell, J.S., "Shape Optimization and Sequential Linear Programming," in *Optimal Structure Design*, Edited by R.H. Gallagher and O.C. Zienkiewicz , John Wiley & Sons, New York, 1973.

[11] Kristensen, E.S. and Madsen, N.F., "On the Optimal Shape of Fillets in Plates Subjected to Multiple In-Plane Loading Cases," *Int. J. Num. Meth. Eng.*, Vol. 10, 1007–1019 (1976).

[12] Braibant, V. and Fleury, C., "Shape Optimal Design Using B-Spline," *Comput. Meth. Appl. Mech. Eng.* Vol. 44, 247–267 (1984).

[13] Naqib, R.A., Zureick, A. and Will, K.M., "New Approximate-Iterative Technique in Shape Optimization of Continuum Structures," *Comput. Struct.*, Vol. 51, 737–748 (1994).

[14] Chong, K.P. and Pinter, W.J., "Stress Concentrations of Tensile Strips with Large Holes," *Comput. Struct.*, Vol. 19(4), 583–589 (1984).

MFPT

OIL ANALYSIS

CoChairs: Thomas H. Crum
C J Analytical Engineering, Inc.

Paul L. Howard
Paul L. Howard Enterprises

CHARACTERIZATION OF DEBRIS FROM F404 ENGINE OIL FILTERS BY ENERGY DISPERSIVE X-RAY FLUORESCENCE

Gary R. Humphrey
Technical Support Center
Joint Oil Analysis Program
296 Farrar Road
Pensacola, FL. 32508-5010
ngrh1@navtap.navy.mil

Abstract: The Department of Defense (DoD) Oil Analysis Program uses rotrode Atomic Emission Spectroscopy (AES) to analyze for elements in lubricants. The determination of the concentrations of elements by AES is limited to particle sizes of approximately 10 microns or less. The current trend in the DoD is to place 3 micron absolute filters in lubricant systems that remove the majority of the particles that AES uses to determine the condition of the equipment rendering AES an unsatisfactory method to monitor these fine-filtered assets.

The F404 engine was removed from the Navy Oil Analysis Program (NOAP) because rotrode AES could not adequately predict the failures that occurred in the engine. The F404 engine oil filter is a 15 micron absolute, 10 micron nominal filter. Slightly coked synthetic oil coats the filter and can effectively trap particles as small as 0.3 microns. Energy Dispersive X-Ray Fluorescence (EDXRF) spectroscopy was chosen to identify the elements in the debris from F404 engines lubricant system filters. The EDXRF methods and procedures developed at the Joint Oil Analysis Program Technical Support Center (JOAP-TSC) can identify abnormal wear in the F404 engineand are used to identify which F404 engine module(s) are undergoing wear and generating wear particles.

Key Words: Element Percentage Index (EPI); Element Thickness Index (ETI); Energy Dispersive X-Ray Fluorescence (EDXRF); F404; Fundamental Parameters (FP)

BACKGROUND: The US Army Oil Analysis Program Office tasked the JOAP-TSC in January 1992 [1] to develop a procedure to analyze the debris trapped by 3 micron filters. A literature review and a survey of commercial, allied nation and DoD laboratories was done to determine the methods in use for identifying and quantifying debris from lubricant system filters. EDXRF was identified from the literature as the most promising technology to use to identify debris from lubricant system filters. The lack of availability of 3 micron filters resulted in the use of F404 engine oil filters for this study. This study firmly established the ability of EDXRF technology to analyze the elements contained in the debris extracted from F404 engine oil filters.

THEORETICAL: The basic atomic mechanism for producing x-ray fluorescence may be understood by considering the interaction of the incoming x-ray with the atomic electrons. Electrons are oriented around elements in energy shells. The electrons of interest in EDXRF analysis are in the K, L and M inner energy shells [2]. The K electrons

are in the energy shell closest to the nucleus and are the most tightly bound electrons in the atom. The electrons in these inner energy levels possess the binding energies necessary for generating the fluorescent x-ray photons measured in EDXRF spectroscopy. An x-ray photon emitted from the source, an x-ray tube, can impinge upon an element's electron and cause the ejection of the electron, leaving behind an orbital vacancy. An electron from another energy shell can move into this vacancy with the emission of an x-ray photon, a process called fluorescence. The emitted x-ray photon carries an energy equal to the difference in the binding energies of the two energy shells, e.g. the more loosely bound outer shell binding energy is subtracted from the binding energy of the more tightly bound inner shell [2]. The value of the emitted x-ray photon's energy is characteristic of the element. This explanation is a simplification of the process and omits other possible interactions, but is quite sufficient for a basic understanding of the relevant processes.

The EDXRF analysis of a large piece, an inch or more in diameter, of metal alloy with a specific composition of elements is straightforward. The percentages reported by an EDXRF spectrometer analysis can be related to the concentration of each element in the alloy. However, the debris from oil filters is composed of metal particles from many different alloys. The percentage of a given element in this mixture of several alloys does not have the same meaning as the percentage of the element in a single alloy and is not directly related to the elemental concentrations in any of the individual particles. The elemental analysis of wear particulates by EDXRF, for the purposes of condition monitoring, can be carried out quite effectively by a statistical analysis of the percentages alone without addressing the theoretically complicated issue of concentration for these particulates.

Intermediate thickness films, large particles, and bulk samples of alloys can exhibit matrix effects, where the fluorescent intensity radiated, per atom of a given element, depends upon the alloy composition. This complication can be handled with a Fundamental Parameters (FP) calculation. The FP calculation produces the matrix correction factors (alphas) and the estimate of sample thickness. The FP calculation estimates the thickness by comparing measured x-ray intensities to theoretical x-ray intensities expected from a sample of a given thickness. The FP calculation goes through iterations until the program finds the thickness which matches the theoretical x-ray intensities with the measured x-ray intensities. Once the thickness is found, the matrix correction factors (alphas) are calculated for the unknown sample [3]. Once the FP analysis is complete, the corrected data may be analyzed and compared with the engine history.

INSTRUMENT: The instrument that was used in this study is a Spectrace 6000 Energy Dispersive X-Ray Fluorescence (EDXRF) spectrometer system manufactured by Spectrace Instruments, Inc. The system is modular. The system is composed of 5 parts or modules: the EDXRF spectrometer, the electronic card cage, a personal computer, a printer and a vacuum pump. The modular design is compatible with JOAP laboratories

that have limited available space, e.g. ships. The modular design also facilitates trouble shooting and replacement of the malfunctioning module.

PROCEDURES: EDXRF samples are prepared from F404 engine oil filters. Protective surgical gloves and eye protection are worn during this procedure. Debris is separated from the oil filter by submerging the oil filter in Electron[4] solvent and applying ultrasonic cleaning for 5 minutes. The debris is then captured on a 1micron absolute, polycarbonate filter using a vacuum to expedite the filtration process. The debris captured by the filter is introduced to the x-ray system. The filtering of the debris takes the most time. One entire cycle takes approximately 20 minutes. However, when doing multiple samples, the cycles overlap and the time is shortened to approximately 15 minutes.

DEVELOPMENT OF ELEMENT GUIDELINES: A set of guidelines was developed to determine which elements and combinations of elements are influential in identifying the source of the wear particulates. The process used to develop the element guidelines was: 1) Obtain the list of alloys that composes each of the oil wetted parts in each module of the F404 engine. Research each alloy composition specified for the oil wetted parts. Define the elements that compose each alloy and the percent composition of each element in the alloy. 2) Analyze the debris captured by the engine oil filter by EDXRF. 3) Develop EDXRF indexes and guidelines. 4) Compare the results to Aircraft Engine Management System (AEMS) reports.

A total of 189 filters from F404 engines were analyzed by EDXRF. The majority of the filters came from NAS Lemoore Aircraft Intermediate Maintenance Depot (AIMD) and some from NAS Cecil Field AIMD.

Two empirical indexes have been developed to characterize the significance of an element's presence in the sample. The Element Percentage Index (EPI) is useful for comparing the relative abundance of elements within a given sample. The Element Thickness Index (ETI) further allows for the thickness of the sample, as obtained from the FP calculation, to be considered. These two index values convey the statistical level of significance of each element, with larger values implying more advanced wear. The EPI and the ETI values give complementary information and are used together to ascertain the condition of the asset.

The EPI for each sample's elements is derived from the EDXRF spectrometer's normalized reported percentages. The more an element's normalized percentage deviates positively from the mean, the higher the level of significance assigned to the element's percentage, e.g. standard deviation of 1, 2, 3, 4 or 5. A level of significance for an element of 1 or less is considered "Normal" wear. A level of significance of 4 or 5 signifies an element's percentage had a very large positive deviation from the mean, suggesting advanced, detrimental wear.

The ETI is calculated in the same fashion as the EPI, but the normalized percentages are first all multiplied by the thickness value obtained from the FP calculation for that filter sample. The ETI was developed to take into account the total amount of material in the sample, while also characterizing the significance of the element's presence in a sample. Again, levels of significance are attached to each element's thickness index. The levels of significance of each element's thickness index are derived by the same process as the EPI levels of significance.

This empirical approach to developing a set of indexes allows for the identification of the elements present and the calculation of statistical levels of significance for each element. The mere presence of an element is not enough to indicate a problem, unless the EPI or ETI for that element is abnormally large.

A shorthand notation was developed for use in tabulating the index values. The EPI notation is composed of the element's chemical symbol and a numerical designation of 2, 3, 4 or 5 followed by a "P", e.g. Ti-5P for titanium with a significance level of 5. The ETI notation has the element's chemical symbol and a numerical designation of 2, 3, 4 or 5 followed by a "T", e.g. Ti-5T.

EDXRF guidelines were developed after a detailed consideration of comparisons between the EDXRF results and the metallurgy of the F404 engine. The EPI and ETI values which satisfy the guidelines are then used to identify the engine modules which are producing the wear particles. The guidelines are as follows: 1) An element is considered in evaluating a sample if the level of significance is 2 or greater. The higher the number the more detrimental the wear is to the engine. 2) Fe, Ni and Ti are the 3 major elements used to evaluate the condition of the F404 engine. 3) Fe, Ni or Ti must be present for Ag, Cd and W to be considered in the evaluation of a sample. 4) Al and Si must be present at the 3, 4 or 5 level of significance to be considered in the evaluation of a sample. 5) Ni must be present for Nb to be considered in the evaluation of the sample. 6) Ti must be present for Sn to be considered in the evaluation of a sample. 7) Fe and/or Ti must be present for V to be considered in the evaluation of a sample. 8) Co and Mo must occur simultaneously to be considered in the evaluation of a sample. This combination of elements, Co and Mo, indicates tribaloy coating. 9) Cu, Pb and Zn must occur simultaneously to be considered in the evaluation of a sample. This combination of elements, Cu, Pb and Zn, indicates wear of the oil pump bearing.

The guidelines were statistically developed and then the strength of the correlation between significance levels and engine wear was assessed. The strength of the correlation is depicted in Fig. 1. Examples are featured in the "Discussion of Results". The comparison between the EDXRF results and AEMS reports is taken up in the "Discussion of Results" section, below.

DISCUSSION OF RESULTS: The debris in the F404 engine oil filter is associated with the Oil Tank, Oil Pump, Accessory Gearbox (AGB), Fan (F), High Pressure Compressor (HPC), High Pressure Turbine (HPT) and Low Pressure Turbine (LPT). The

AEMS only details problems with the F, HPC, HPT and LPT. The information gathered from EDXRF spectrometer measurements coupled with the metallurgy of the engine oil system demonstrates that the Oil Tank, Oil Pump and AGB contribute significant debris to the engine oil system. The majority of the alloys in the Oil Tank and Oil Pump are aluminum alloys. The Oil Tank, Oil Pump, AGB and F are the major contributors of Aluminum (Al) to the F404 engine oil system. The presence of water in a polyol ester, e.g. MIL-L-23699 used to lubricate the F404 engine, provides the ingredients that cause corrosion. Acids are formed by the breakdown of the oil into the reactants that originally formed the oil, e.g. organic acids and an alcohol. Al is very susceptible to corrosion by acids. Corrosion induced by the intrusion of water into the F404 engine oil system has been attributed as one of the mechanisms to induce a bearing failure in the accessory gearbox [5]. This is a major concern in this engine. The possibility of Al being generated by corrosion resulted in the level of significance for Al being a 3, 4 or 5 to be considered in an evaluation. However, damage to a housing must be considered at a level of significance for Al of 4 or 5. Silicon is considered relevant to the evaluation at a significance level of 3, 4 or 5 and indicates dirt contamination. Both of these instances for Al and Si are based upon the author's experience in the field of oil analysis. The Oil Pump has a unique alloy composed primarily of copper, lead and zinc. The occurrence of these three metals in a sample indicates wear of the oil pump bearing.

The close correlation between the AEMS history and EDXRF results is exemplified by several specific examples. Several categories of abnormal EDXRF results and the maintenance performed on the engine are detailed to demonstrate the correlation. The categories are as follows: 1) abnormal wear detected in module(s) prior to engine failures, 2) abnormal wear detected in module(s) coinciding with engine failures, 3) Foreign Object Damage (FOD) and 4) unusually high levels of dirt contamination.

In this study, 27 oil filters representing 24 engines experienced failures of engine modules associated with the engine oil system. Twelve oil filters were removed prior to engine failure. The most important characteristic of this EDXRF technique is its capability to detect and associate significant levels of elements with the engine modules that subsequently failed. In each of the 12 cases, this EDXRF analysis found significant levels of the elements that could be associated with the modules that failed in the engines. Two cases are presented in detail for each category -- engine oil filters taken prior to engine failure, engine oil filters taken at engine failure, engine oil filters taken at Foreign Object Damage (FOD) failure and filters taken exhibiting dirt contamination.

FILTERS TAKEN PRIOR TO ENGINE FAILURE: Samples 75 and 83 are from engine serial number 310656. Sample 75 engine oil filter was removed at 2418 Hours Since New (HSN). The elements of concern and their respective levels of significance are Ti-4P, Ti-2T, V-4P, V-2T, W-2P and Al-3P. This combination of metals -- Ti, V and W -- can originate from the HPC. Sample 83 engine oil filter was removed at 2515 HSN. The elements of concern and their respective levels of significance are Ti-4P, V-4P, Sn-4P and Al-5P. Sample 83 indicates a large and abnormal amounts of Ti, V and Sn are being generated. The Ti, V and Sn indicate the wearing of titanium alloys. This

combination of metals --Ti, V and Sn -- can originate from the Fan and HPC. The abnormal amount of Al indicates a housing could be involved or severe corrosion exists. At 2660 HSN, engine serial number 310656 experienced an HPC failure. The first sample at 2418 HSN indicated the HPC was the source of the elements. The subsequent sample also confirmed the HPC could be the source of the elements. This EDXRF method indicated the HPC was the source of the elevated levels of wear debris at 142 operating hours before the HPC failed.

Samples 68 and 189 are from engine serial number 310810. Sample 68 engine oil filter was removed at 2548 HSN. The elements of concern and their respective levels of significance are Ti-2P, V-2P, Co-3P, Co-2T and Mo-2P. The Ti and V indicate a Ti 6-4 alloy and the Co and Mo indicate tribaloy coating. This combination of metals -- Ti, V, Co and Mo -- can originate from the F and HPC. Sample 189 engine oil filter was removed at 2787 HSN. The elements of concern and their respective levels of significance are Fe-5P, Fe-2T, V-4P, V-3T and Cd-2T. The Fe and V indicates abnormal bearing wear with a significant level of Cd in the sample. This combination of metals -- Fe, V and Cd -- can originate from the HPC and LPT. Notice the radical change in the elements and their significance levels from the sample taken at 2548 HSN and 2787 HSN. At 2800 HSN, this engine experienced a compressor failure. The elements in both samples indicated the HPC was a source of these elements. This EDXRF technique indicated a problem at 252 operating hours before the HPC failed. At 2787 HSN or 13 operating hours prior to the failure, this EDXRF technique indicated a bearing failure was occurring.

FILTERS TAKEN AT ENGINE FAILURE: This set of examples illustrates the ability of this EDXRF technique to detect significant levels of elements that can be associated with the engine modules that failed. Fifteen engine oil filters coincided with the failure of engine modules. In every case, this EDXRF technique detected significant levels of elements that can be associated with the engine modules that failed. Sample 13 is from engine serial number 311262. The engine oil filter was removed at 2022 HSN. The elements of concern and their respective levels of significance are Fe-3P, Fe-3T, Ag-3P, Ag-2T, W-3P, W-3T and Cd-2T. The Fe indicates abnormal bearing wear, Ag plating is being generated and a significant level of W and Cd in the sample. This combination of metals -- Fe, W, Cd and Ag -- can originate from the HPC and LPT modules. The AEMS report for sample 13 reported a main turbine failure at 2022 HSN. The F, HPC, HPT and LPT modules were replaced.

Sample 93 is from engine serial number 310075. The engine oil filter was removed at 3536 HSN. The elements of concern and their respective levels of significance are Ti-5P, Ti-4T, V-5P and V-4T. This is a very large and abnormal amount of these 2 elements and is a strong indication of the presence of Ti 6-4 alloy. This combination of metals can originate from the AGB, F and HPC modules. This indicates abnormal wear or a failure has occurred. The AEMS report for sample 93 stated the Fan module had a cracked case at 3536 HSN. The Fan case is composed of Ti 6-4 alloy.

FILTERS TAKEN AT ENGINE FOD FAILURE: Foreign Object Damage (FOD) caused approximately 25% of the maintenance on the F404 engines in this study. There were 39 cases of FOD failure and this EDXRF technique detected significant levels of elements in 33 of the cases with 22 of the cases having significant levels of the three major elements, Fe, Ni and Ti. A FOD event may or may not adversely impact the engine oil system. The important concept is the EDXRF technique must always detect engine oil system failures. The EDXRF technique did this in all of the cases in the "Filters Taken Prior To Engine Failure" and "Filters Taken At Engine Failure" categories above. There was a wide variety of elements generated by FOD intervention with indications that F404 modules were seriously impacted. Sample number 1 is from engine serial number 360018. The engine oil filter was removed at 583 HSN. The elements of importance and their respective levels of significance are Nb-3P, Nb-5T, Ag-2P, Ag-3T, Fe-3T, Ni-3T, W-2P and W-2T. The Ni and Nb indicate abnormal wear is originating from a Ni alloy, the Fe indicates abnormal bearing wear with Ag plating and W being generated. This combination of metals -- Fe, Ni, Nb, W and Ag -- indicates the HPC, HPT and LPT modules can be the sources of the metals. According to the AEMS report, this engine had maintenance performed for FOD at 737 HSN. The Fan, HPC, HPT and LPT were replaced for FOD.

Sample 18 is from engine serial number 311465. The engine oil filter was removed at 536 HSN. The elements of concern and their respective levels of significance are Ni-2P, Ni-3T, Nb-3P, Nb-5T, Fe-3T, W-2P and W-3T. The Ni and Nb indicate abnormal wear is originating from a Ni alloy; the Fe indicates abnormal bearing wear with a significant level of W in the sample. This combination of metals -- Ni, Fe, Nb and W -- can originate from the HPC, HPT and LPT. According to the AEMS report, this engine had maintenance performed for engine FOD in the HPC at 536 HSN. The HPC and HPT were replaced.

FILTERS EXHIBITING DIRT CONTAMINATION: The capability of EDXRF to detect Silicon (Si), combined with these procedures, is another example of the ability of EDXRF to detect oil system contaminants. Si at a significance level of 3, 4 or 5 is an indication of dirt contamination in the oil system. Sample 17 is from engine serial number 310819. The engine oil filter was removed at 2110 HSN. The element of importance and level of significance is; Si-3P and Si-5T. The F, HPC, HPT and LPT were replaced for a high time fuel nozzle.

Sample number 49 is from engine serial number 311252. The engine oil filter was removed at 2189 HSN. The element of importance and level of significance is Si-5P. The HPC and Fan were replaced for FOD and the HPT was also replaced at 2189 HSN.

EDXRF CORRELATIONS WITH AEMS: Correlations between EDXRF and the AEMS reports are discussed in this section by referring to the chart. In the chart below, the major element with the highest level of significance for the oil filter is accounted for in the respective AEMS maintenance category. This chart includes the 4 major categories depicted in the AEMS reports. The chart depicts several important aspects of monitoring

the condition of the F404 engine. A significance level of 2 occurs 5 times for the 3 major elements with no maintenance reported by AEMS and is the category labeled "Normal". The "Normal" category represents AEMS reports on engines that are operating normally or no maintenance was performed on the engine. The significance level of 2 occurs in each category, including engine failure. In assessing the condition of F404 engines by EDXRF, a 2 level of significance for the 3 major elements is important because of the frequency this level occurs in the engine "Failure" category. The significance levels of 3, 4 and 5 do not occur in the "Normal" category and substantiate the observation that increased wear is occurring and is detrimental to the engine. A level of significance of 3 is found in the "FOD", "High Time" and "Failure" categories. Again, the frequency of the 3 level of significance within the engine "Failure" category is important in evaluating the condition of the engine. A level of significance of 4 represents a value that has exceeded the recalculated mean by more than 3 (recalculated) standard deviations.

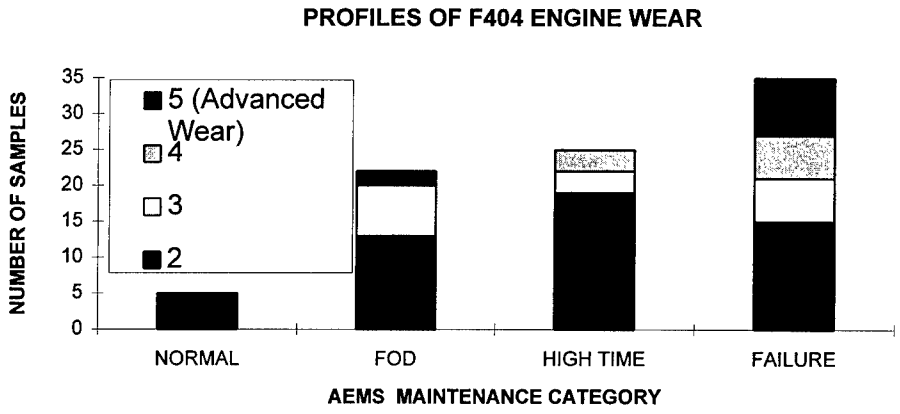


Figure 1. The levels of significance for the three major elements, and their occurrence in each AEMS engine maintenance category.

A rating in this significance level is found on the chart in the "High Time" category and mostly in the engine "Failure" category. The 5 level of significance represents wear that is the most detrimental to the engine. This rating almost exclusively occurs in the engine failure category (also in FOD cases) and is considered a strong indicator that a failure is in progress and/or has occurred. As stated previously, any one of the significance levels, e.g. 2, 3, 4, or 5 can indicate an engine failure. A significance level of 5 virtually ensures an engine failure has occurred or will occur.

Foreign Object Damage (FOD) accelerates the wear in mechanical systems. FOD also generates wear that is detected by EDXRF. FOD generates a wide variety of elements and element levels of significance, up to a 5 level of significance, in some cases, indicating severe damage.

The removal of engine modules for "High Time" is also depicted in the chart. There were 61 filters removed that coincided with "High Time" maintenance. "High Time" refers to the removal of an engine module after a specific number of hours. The removal of the engine module does not depend upon whether the wear it has incurred during its cycle of operation has debilitated the module to a point where the module must be removed. The removal of the module is based solely on the time the module has been in operation. This scheduled maintenance philosophy is used on the F404 engine. This chart illustrates that in 25 cases, this EDXRF technique agreed that the modules were generating elevated levels of wear of the 3 major elements, Fe, Ni and Ti, when maintenance was performed.

The primary reason for evaluating EDXRF technology was to ascertain if the EDXRF results could detect failing F404 engines. The "Failure" category includes oil filters from engines that failed, oil filters from engines that the combustor failed and oil filters from engines that the afterburner failed. There were 11 filters associated with afterburner failures with 5 of the afterburner failures coinciding with the generation of the major EDXRF elements, Fe, Ni and Ti. There were 6 combustor failures with 3 of the combustor failures coinciding with the generation of the major EDXRF elements. An afterburner or combustor failure may or may not have an impact on the engine oil system. The afterburner and combustor are not involved with the engine oil system however, major EDXRF elements were generated up to a 5 level of significance with these failures. The important aspect is the 27 oil filters associated with the engine oil system had major EDXRF elements present at a significance level of 2 or greater in each analysis. This definitively shows a direct correlation with F404 engine failure and the interpretations of the EDXRF analysis previously discussed. Of the 27 engine oil filters, 12 of the engine oil filters were removed prior to the engine failing. In each case, the major EDXRF elements were present at a significance level of 2 or greater. This demonstrates conclusively the ability of these EDXRF techniques to predict engine failure. The EDXRF techniques also eliminates modules the elements can originate from and includes the modules that actually failed. This will assist greatly in performing maintenance on the F404 engine.

In the field of oil analysis, a diagnostic method must be able to successfully diagnose and predict when engines will fail. The engine is removed from service prior to failing. Any other predictive property the procedure has is an added feature. This EDXRF method can diagnose F404 engines that will fail. EDXRF technology is capable of indicating the presence of abnormal amounts of elements in the F404 engine oil system and indicating the F404 module(s) that are generating those elements. The examples presented in detail in this discussion section are a small representation of the data.

DIFFERENCES BETWEEN EDXRF AND AEMS: Whereas the previous section presented correlations between EDXRF and AEMS, this section discusses cases where they do not correlate.

There were 36 filters where the AEMS record indicated normal operation. In 31 of these cases, this EDXRF technique did not detect significant levels of the major elements Fe,

Ni and Ti. In 5 instances this EDXRF technique indicates a problem and the AEMS data does not confirm this. For example, samples 58 and 59 are from engine serial number 360039. Sample 58 engine oil filter had abnormal levels of nickel alloy, however the AEMS record indicated "normal operation". Sample 59 engine oil filter was taken 5 hours later and the EDXRF technique indicated the sample was normal and the AEMS record depicted "normal operation". The reason for pulling the engine oil filter after only 5 hours is unexplained. The other 4 instances are unconfirmed and possibly could have been explained if additional engine oil filters from their respective engines had been submitted for analysis. The submission of consecutive F404 engine oil filters from the same engine by maintenance personnel was not a requirement. This is why the other 4 instances could not be confirmed. The study should be expanded to elucidate the differences.

The AEMS database cannot confirm EDXRF indications that point to the oil tank, oil pump or AGB because these components are not detailed in the AEMS database.

In several of the failed engines, EDXRF reported fewer modules could be the sources of the elements than were actually replaced. It is difficult to determine from the AEMS database whether or not secondary damage accounted for the additional failed modules. The EDXRF analysis identified and included the module that failed. In the future, engine oil filters from the same engine would be analyzed by EDXRF on a regular basis. This will improve the capability of the EDXRF procedure to identify the failing module. The maintenance information entered into the AEMS system on F404 modules must also supply information about the oil tank, oil pump and accessory gearbox.

CONCLUSIONS: EDXRF is a nondestructive technology and provides valuable information on the condition of the F404 engine. The engine condition information provided by EDXRF, using the methods outlined in this report, correlate very well with the maintenance performed on the F404 modules contained in the AEMS report. This procedure can monitor wear in the F404 Oil Tank, Oil Pump, AGB, Fan, HPC, HPT and LPT. Fifteen engines suffered failures during this study and, in each case, the EDXRF technique detected significant levels of the elements and included the engine module that later failed. More importantly, a third of the failed engines had oil filters submitted prior to failure. The module that subsequently failed was included in the modules indicated by EDXRF. The EDXRF data provided 100 or more hours of advanced warning of engine failure.

The procedure can be used by field personnel to predict and prevent serious damage to the F404 engine. This part of the project has not included a field personnel evaluation. This EDXRF decision scheme can be adapted to the existing US Army and Navy oil analysis software, Oil Analysis Standard Interservice System (OASIS). This will provide an evaluation platform already familiar to field personnel. The cost will be minimal to produce the software code because the JOAP-TSC has written the software code for the US Navy module of the OASIS program and is a custodian of the code for the OASIS software.

BIBLIOGRAPHY

1. Peter Fast to Director, Joint Oil Analysis Program Technical Support Center, 3 January 1992, *Analysis of Wear-metal Debris from Oil Filters.*

2. R. Jenkins, R. W. Gould, D. Gedcke, *Quantitative X-Ray Spectrometry, 2nd Ed.,* (Marcel Dekker, Inc., 1995).

3. Spectrace Instruments, Inc., 345 East Middlefield Road, Mountain View, California 940343, *Fundamental Parameters, Version 1.34*, (January 1991), 6-1.

4. Sentry Chemical Company, 1481 Rock Mountain Blvd., Stone Mountain, Ga. 30086.

5. Conversation with Martha Irene, F404 Engineer.

REAL TIME OPTICAL OIL DEBRIS MONITORS

J. Reintjes, R. Mahon*, M. D. Duncan, L. L. Tankersley†, J. E. Tucker, A. Schultz**, V. C. Chen**, C. Lu**, T. L. McClelland, P. L. Howard***, S. Raghavan††, and C. L. Stevens††

Laser Physics Branch, US Naval Research Laboratory
Washington, DC 20375

Abstract: The performance characteristics of two optical debris monitors are described. The monitors allow real-time, on-line identification of the type and severity of faults in machinery, through measurement of size, shape and morphology of debris particles.

Key Words: Bearings; early warning; catastrophic failure; gears; hydraulic fluid; real-time; shape classification; wear debris

Introduction: We have previously described the development of real-time, imaging, optical debris monitors. [1,2,3,4]. These allow real-time, or near real-time, detection of ferrous, non-ferrous and non metallic debris in lubricating oil or other fluids in the general size range above about 5 micrometers. The optical debris monitors provide the capability for determining the type and severity of faults through combined analysis of particle size, rate of production and particle morphology [5-8]. One major advantage of this approach to debris monitoring is the capability of on-line identification of specific types of fault, such as cutting or fatigue spalling. A second advantage is the ability of the optical monitors to detect relatively small particles. These properties allow the independent progression of specific faults to be followed from an early stage, improving diagnostics for condition based maintenance, and providing improved early warning of catastrophic failures.

The general form of the optical debris monitor is illustrated in Fig. 1. It consists of a laser illuminator, an image detector and an image processor. Because of limitations associated with optical resolution and depth of field we have found it convenient to break the size of the detected particles into two ranges : 1) larger than about 50 μm and 2) larger than about 5 μm.

The large particle detector, referred to as LASERNET, is appropriate for detecting failure-related debris in machinery lubricating oil in real time. One application of the monitor in this form is in helicopters to provide improved early warning of faults before they advance to catastrophic failures. Because the relatively large particles are trapped in the filter, the LASERNET detector will typically be located in the scavenge oil line, ahead of the filter. The particles must be detected on their only pass through the viewing cell. Typical operating conditions (flow speeds of 10 m/sec, viewing apertures of ~ 1 cm), along with the requirement of illuminating all of the oil column to maximize the detection efficiency, requires that the laser pulse duration be of the order of 1-2 μsec, the repetition rate be of the order of 0.5 - 1 kHz and that the image detector have approximately 1000 x 1000 elements and be capable of framing at 1 kHz. The images must be scanned in real time to determine the presence and characteristics of debris particles.

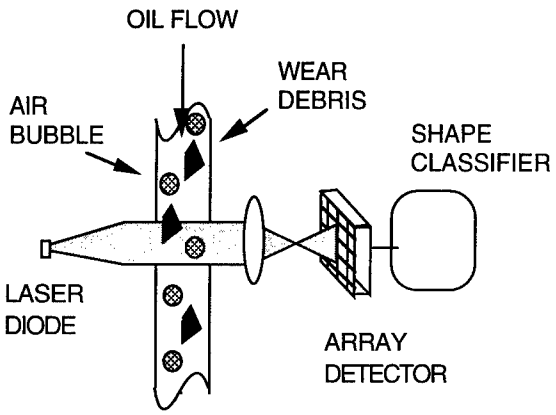
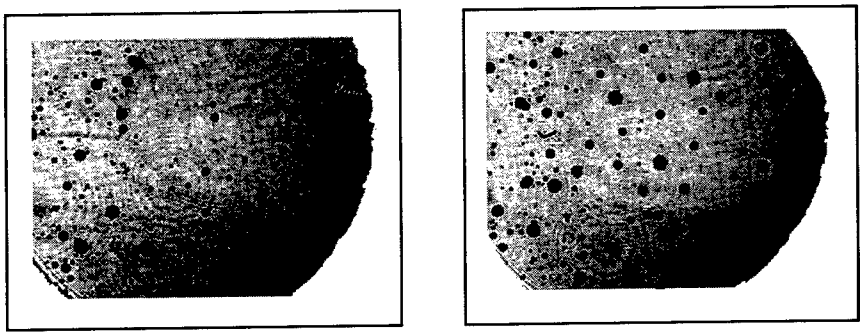


Fig. 1. A schematic diagram of the optical oil debris monitor, showing the illuminating laser diode, the array photodetector and the shape classifier. Objects suspended in the oil flow are imaged in transmission onto the array detector, appearing as dark shadows against a bright background.

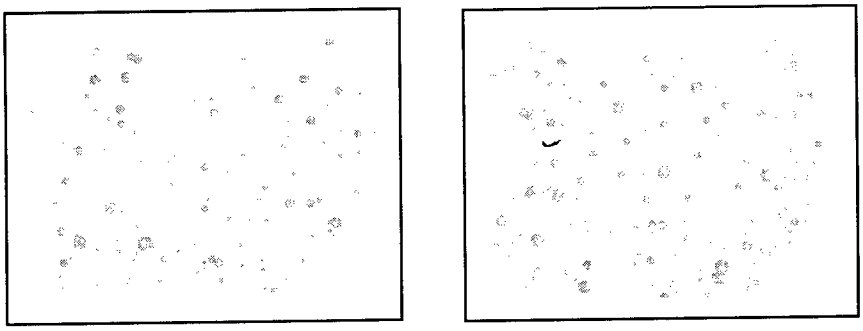
LASERNET AVDS SYSTEM: We have implemented the LASERNET detector for test on the T700 engine and the main gearbox of an SH-60 power train in the test cell at the Naval Air Warfare Center Aircraft Division at Trenton, NJ under the US Navy's Air Vehicle Diagnostic System (AVDS) Advanced Technology Demonstration program. The test system has been described previously [3,4]. Briefly it consists of a laser illuminator, a flow cell adapter, a high speed camera and a high speed image processing system. For the tests at Trenton, additional components for test validation have been included. These consist of a video camera for visualizing the flow conditions, an auxiliary high speed memory and a standard chip detector.

The illuminator is a commercial laser diode operating at 830 nm. The laser is coupled to a single mode fiber and the output beam expands to the desired final diameter and is collimated with an output lens. The final diameter is chosen to be larger than the illuminated area of the LASERNET viewing cell to ensure uniformity of illumination for the image detector. Typical center to edge variation in intensity is 0.8. The center of the viewing cell is imaged onto a high-speed camera from Silicon Mountain Design. This camera is capable of framing at speeds up to 1 kHz with up to 8 bits per pixel. In order to accommodate the speed requirements of the image analysis in the LASERNET system we have implemented the camera with 1 bit for the current set of tests. The high speed camera thus provides black and white images for analysis. Some control over the details of the image can be obtained by adjusting the bright-dark threshold of the camera and the peak laser intensity. Each existing engine or gearbox requires its own adapter and fittings for retrofit. For the engine tests at the Trenton test cell we have chosen to attach the LASERNET system to the engine auxiliary gearbox manifold at the chip detector port.



a b

Figure 2. Images of typical flow conditions on the T700 test engine without debris (a) and with an occasional debris particle (b).



a b

Figure 3. Results of processing images in Fig. 2 with high speed image processor. Light objects are identified as air bubbles, while dark object is identified as debris particle.

The major challenge in the development of the LASERNET system are detection of debris and rejection of air bubbles at the speed necessary to accommodate the full oil flow and with a low false alarm rate. In our system air bubbles are identified on the basis of their round shape, while debris particles larger than 50 μm will have irregular shapes. For the specific flow conditions of the T-700 engine the laser/camera system is operated at 500 frames per second. Early operation indicated the ability of the system to detect and identify debris. An example of fabrication debris seen during the initial operation is shown in Figs. 2 and 3. In Fig. 2 two raw images of the flow are shown, one with a piece of debris and one without. The same images are shown in Fig. 3 with the debris correctly identified by the image processor. The system is currently undergoing testing on the T700, which is expected to be a relatively clean system, to verify an acceptably low false alarm rate.

LASERNET FINES: The small particle detector, LASERNET FINES, is designed to detect and classify particles in the size range above 5 μm. It is appropriate for detection of debris particles associated with normal and early wear in lubricating systems. It is also capable of detecting particulate contamination in hydraulic and fuel systems, and in other fluids. Again, the goal is to identify type and severity of fault through determination of particle size distribution and morphology. However, since it is sufficient to determine the representative concentration of particles in the size range of LASERNET FINES, rather than counting each and every particle, sampling in time and space is acceptable. This allows a reduction in the amount of fluid sampled and slower data detection and processing rates that are compatible with standard tv formats, reducing the special demands on the imaging system. The depth-of-field requirements for detection of the smaller particles requires use of cells that are of the order of 100-200 μm thick. The increased spatial resolution required for detection of the smaller particles is obtained with magnification of the image of the order of 6x. With the particle concentrations that are typically encountered, particle distributions from different frames must be combined to obtain statistically significant results. Total detection time and image processing rates can be traded within an overall system. The LASERNET fines system can be configured as a batch processor, a full autonomous on-line processor or a hybrid in which the sampler is located in-line, but the fluid sample is ported to the unit.

We have currently configured the LASERNET Fines unit as a benchtop batch processor. In the present unit the illuminator is a commercial laser diode operating at 830 nm with 100 mW peak power. The fluid is gravity-fed through a cell approximately 100 μm thick. In our current configuration the image is obtained with a Cohu full-field tv camera using 6x magnification. The images are captured and processed on a Power Macintosh 8100 computer.

The image processing code for the LASERNET FINES is different from that for LASERNET in several respects. First, round objects cannot be discarded from consideration because there are wear sources that can produce round debris particles in this size range, such as particles from fatigue cracks. Secondly, more than one particle is expected in each frame. As a result, the program does not test objects for roundness and discard round objects, but rather counts and analyzes all objects in the frame. However, because of the sample handling methods, it is anticipated that air bubbles will be much less of a problem. In the event that they remain, side illumination can be used to give the bubbles a characteristic highlight as discussed in previous papers.

The LASERNET FINES system has been exercised with a variety of fluids including contaminated hydraulic fluid, calibration hydraulic samples, using AC fine test dust, and diesel lubrication oil. We have previously reported on the first two [3,4]. A sample of the images obtained with the diesel sample is shown in Fig. 4 along with an image from hydraulic fluid. Although the oil is very black to the eye, the laser diode can image through a path length of about 50-100 micrometers without trouble.

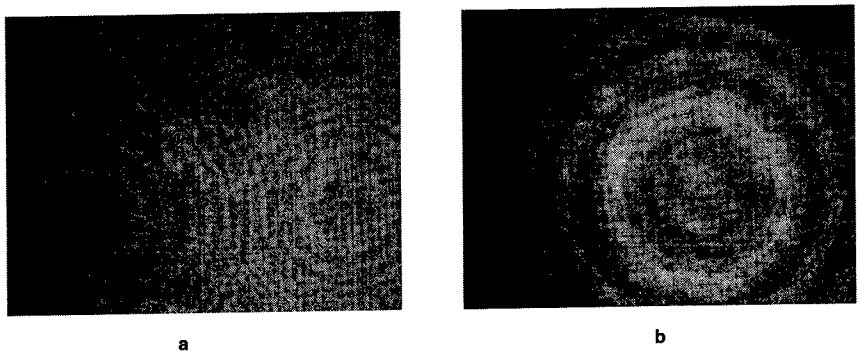


Figure 4. Typical image through hydraulic fluid in 100 micrometer cell (a) and diesel lubricant in 50 micrometer cell (b) taken with LASERNET FINES system.

Future development will include on-line operation of the LASERNET FINES system to provide reduced manpower operation and full automation of oil debris monitoring with detection of debris particles down to the 5 micrometer level. Realization of on-line operation requires production of a suitable sample generator that can provide sample viewing paths of the order of 100 micrometers while providing sufficient pressure for operation along the lubrication system. Combination with FTIR spectroscopic systems will provide a comprehensive, optically-based, oil-analysis system that provides both oil condition analysis and debris detection.

*Jaycor, Vienna, VA
†Dept. of Physics, US Naval Academy, Annapolis, MD
**Naval Research Laboratory, Code 5362, Washington, DC 20375
***P. L. Howard Enterprises, 1212 Clearbrook Rd., West Chester, PA
††LNK Corp., 6811 Kenilworth Ave., Suite 306, Riverdale, MD 20737

References

1. J. Reintjes, R. Mahon, M. D. Duncan, L. L. Tankersley, A. Schultz, V. C. Chen, D. J. Kover, P. L. Howard, M. Chamberlain, Srini Raghavan, and Naresh Gupta, "Optical Debris Monitoring", JOAP Annual Meeting, Pensacola FLA, November 1994
2. J. Reintjes, R. Mahon, M. D. Duncan, L. L. Tankersley, A. Schultz, V. C. Chen, D. J. Kover, P. L. Howard, M. Chamberlain, Srini Raghavan, and Naresh Gupta "Optical Oil Debris Monitor", in "Life Extension of Aging Machinery and Structures", H. C. Pusey and S. Pusey, eds. pp. 57-66, 1994

3. J. Reintjes, R. Mahon, M. D. Duncan, L. L. Tankersley, A. Schultz, V. C. Chen, P. L. Howard, Srini Raghavan, and Naresh Gupta, "Advances in optical oil debris monitoring technology", Integrated Monitoring, Diagnostics and Failure Prevention, MFPT Society, H. C. Pusey and S. Pusey, eds. pp. 269-276, 1996
4. J. Reintjes, R. Mahon, M. D. Duncan, L. L. Tankersley, T. L. McClelland, A. Schultz, V. C. Chen, P. L. Howard, Srini Raghavan, and Naresh Gupta, "Real Time Optical Debris Monitoring", in "Monitoring Technology for Condition Based Maintenance, International Meeting of ASME, 1996.
5. A. Albidewi., A. R. Luxmore, B. J. Roylance, and G. Wang, "Determination of Particle Shape by Image Analysis-the Basis for Developing an Expert System," in "Condition Monitoring '91," M. H. Jones, J. Guttenberger and H. Brenneke, eds., Pineridge Press, Swansea, UK, 1991, p. 411
6. B. J. Roylance and S. Raadnui, "The morphological attributes of wear particles - their role in identifying wear mechanisms", Wear **175**, 115 (1994).
7. B. J. Roylance, I. A. Albidewi, M. S. Laghari, A. R. Luxmore and F. Deravi, "Computer-Aided Vision Engineering (CAVE) - Quantification of Wear Particle Morphology", Lubr. Eng. **50**, 111 (1993)
8. J. J. Hamalainen and P. Enwald " Inspection of wear particles in oils by using a fuzzy classifier", SPIE vol 2249 "Automated 3D and 2D Vision", 390 (1994).

FILTER DEBRIS ANALYSIS USING XRF

Robert R. Whitlock
Code 6680
Naval Research Laboratory
4555 Overlook Avenue SW
Washington DC 20375
robert.whitlock@nrl.navy.mil

Abstract: Monitoring the wear condition of oil-lubricated machinery has traditionally been performed by atomic emission spectroscopy, despite the technique's emphasis on fine particles and its inability to measure the >10 micrometer (>0.0004 inch) particles characteristic of abnormal wear. A growing body of machine types is being serviced with fine filters to remove the fine wear particulates from the circulating oil, leaving none for traditional oil analysis to measure. Benefits of fine filtration are gained in machine life, but failure warning is thereby lost. Warning and condition monitoring functions may be restored with the aid of bench top or portable X-Ray Fluorescence (XRF) instrumentation for analyzing the wear debris collected as a thin layer on membrane filters. For coarsely filtered machines, a drawn oil sample of a few milliliters is membrane-filtered and x-ray analyzed. For finely filtered machines, the wear particulates are first recovered from the machine's in-line filter, for example by Humphrey's straightforward and rapid technique, and then refiltered onto a membrane. Thus, the same bench top x-ray instrument can serve multiple machines with varying filtration specifications. The x-ray measurements indicate wear rates for each of the over two dozen elements of interest and are useful for identifying worn engine modules and components, while providing as much as 100 hours advanced warning of impending failure and affording new opportunities for management of operations and maintenance.

Keywords: Condition monitoring, elemental analysis, prognostics, wear debris, wear debris, wear particle, x-ray fluorescence analysis, XRF

INTRODUCTION: As the components of machinery move in contact with one another, natural wear processes take place which result in the removal of wear material from the contact surfaces. For systems with circulating lubrication systems, this free wear material may be transported in the lubricant. Wear particle analysis is concerned with analyzing this material to determine the wear condition of the machine. Knowledge of the actual machine condition is important for effective management of expensive assets.[1] The larger field of oil analysis, which includes analysis of suspended wear particles, further evaluates the lubricant for its condition and for the presence of other contaminants, such as fuel, coolant and water. Vibrational analysis and wear particle analysis are related: vibrational analysis measures the acoustic emissions produced by the machine while or after discharging the wear particulates, whereas wear particle analysis measures the mass (and its various attributes) emitted by the wear process. Particles may also be produced by corrosion and by

contamination from sources outside the machine, such as dirt; these contaminant particles also fall in the domain of wear particle analysis.

The ultimate goal of any machinery condition diagnostic, including wear particle analysis, is to answer the key questions, "*Does it* need fixing or attention," and "*Where does it* need fixing or attention?" Wear particle analysis can address the following measurable quantities, which in turn may be evaluated to determine the answer to these key questions: quantity, sizing (size distribution and height, which together provide quantity information), composition, particle morphology,[2] and rates for the previous items. Composition of particles is the most direct indication of the identity of sources of the wear particles. Wear modes are deduced largely from particle morphology. Production rates are derived from measurements carried out at known intervals of machine operating time, while taking into account such factors as concentration in the lubricant, volume of oil, and particle removal rates where applicable. The rate of emission of wear particles from a particular source is the most direct indication of the extent of wear and need for maintenance of the source. *Thus, wear rates and debris composition are the measurables closest to the desired information on wear degree and wear sources.*

COUNTING AND SIZING DIAGNOSTICS: The simplest particle counting device is the magnetic chip detector, which seeks to collect any large metal particle which may be present due to a bearing failure requiring immediate action. Chip detectors are ineffective for particles of size $\leq$40 μm. Various other instruments are available for counting particles and particle size distributions down into the few μm range, typically by analyzing a drawn oil sample. On-line sensing systems are available which can count and even size particles as they pass by in the circulating lubricant. The on-line MetalSCAN sensor, commercially available from GasTOPS, measures magnetic and conductive particles $\geq$100 μm and classifies them by size. The combination of counting and sizing allows the total quantity (number and estimated mass) of measured particulates to be determined with the MetalSCAN. The developmental LASERNET on-line sensor employs laser optical imaging and image processing to count and size particles larger than 50 μm (or 5 μm for the LASERNET fines sensor, also in development). The on-line sensors can be expected to have limitations placed on the flow rates and oil line sizes which they can monitor. Under favorable conditions, monitoring of the full oil flow may be achievable.

COMPOSITIONAL DIAGNOSTICS: Compositional diagnostics are important for identifying wear sources. Ferrography is capable of determining certain compositions by way of visual observations in a microscope. The most fundamental compositional methods employ wet chemical methods, by which the wear debris are separated from the lubricant and analyzed by well established quantitative chemical procedures. Acid digestion of the metals may be used as a prelude to subsequent elemental analysis.

There are several widely useful compositional methods based on atomic spectroscopy. Atomic absorption spectroscopy (AAS) and atomic emission spectroscopy (AES) have both been extensively used for analysis of suspended wear particulates. The various chemical elements absorb and emit light at wavelengths characteristic of the energy levels of the outer

electron orbits. The fingerprint of wavelengths is different for each chemical element. These unique identifying features are used to perform elemental analysis of small quantities of fluid and wear particulates. Though widely useful, the methods are not without complications. There are interelement effects by which the presence of one element can influence the result for another. The physical properties of the fluid can also be important to the sample handling as performed in the instrument. These issues are generally controlled by stringent reference calibration and sample preparation procedures. But foremost is the inability of these methods to determine the composition of particles greater than ~10 μm or even less. Since abnormal wear is known to generate the bulk of the discharged material at the larger particle sizes, this limitation can become crucial.

Nuclear spectroscopy has also been carried out in conjunction with wear particle elemental analysis. The two major approaches here are (1) the irradiation of the original parts, followed by analysis of the wear debris[3] and/or original parts (even as they remain in operation[4], [5]) by their nuclear gamma ray emissions,and (2) collection of normal wear debris which are then activated in a reactor for quantitative analysis.[6] These methods can provide impressive results for design studies and individual investigations. Due to safety, economic, and technical considerations, nuclear methods are not suitable for routine monitoring.

X-RAY COMPOSITIONAL DIAGNOSTICS: Atomic spectroscopy may also be performed by using the inner electrons of the atom whose orbits are close to the nucleus. The photon wavelengths associated with these inner orbits are characteristic of the element, but, instead of producing visible light photons as from outer orbits, these transitions yield x-ray photons with much higher energies. The emitted x-rays may be detected and analyzed to determine which elements are present and in what quantities. Thus, the compositional information gained from AAS and AES for wear metals may also be determined by x-ray methods. The x-ray elemental analysis methods are distinguished from one another by the means used to excite the x-ray emission in the sample. The most common method of x-ray production is by electron bombardment, as occurs in an x-ray tube itself. X-ray fluorescence (XRF) involves radiating the sample with primary x-rays, which stimulate the sample to emit its own characteristic x-rays which then are analyzed. Proton induced x-ray emission (PIXE) analysis may also be performed, but requires a higher investment in equipment. All three methods have been applied to wear particulate analysis.[7] X-ray methods are also classified according to the technique used for determining spectral wavelength or, equivalently, photon energy (an inverse function of wavelength alone). The distinction is between diffraction methods analogous to a prism diffracting visible light (wavelength dispersive x-ray analysis or WDX) versus measuring the energies of electrons produced when x-rays are absorbed in a detector (energy dispersive x-ray analysis or EDX). Of the two, EDX is much more compact, while WDX offers the best spectral resolution.

The composition of individual wear particles is readily determined using x-ray emission measurements carried out when the sample is illuminated by the electron beam of a scanning electron microscope (SEM) or electron microprobe. The SEM x-ray methods are very convenient for single particle analysis, because the particle may be imaged while being locally analyzed to determine its constituent chemical elements. However, this method is laborious

if applied to very many particles and, in general, is not economical as a routine monitoring method. Ferrographic analysis, in addition to its use of the visual microscope for identifying oxides, is typically augmented with SEM x-ray measurements to determine wear metal composition.

The batch processing of an entire sample of wear particulates is conveniently performed using elemental analysis by x-ray fluorescence (XRF). Both XRF and the SEM use the same characteristic x-ray wavelengths to identify the elemental constituents of the sample wear metals, from which wear sources may be identified. XRF also determines the quantity of each element present in the sample and is suitable for measuring overall ferrous wear rates, while simultaneously determining rates for compositionally distinctive sources. *Thus, XRF determines both wear rates and composition, the two most direct indicators of the need for maintenance and the source of wear, which are the key issues for any machinery condition monitor.*

Because XRF does not involve direct irradiation of the sample by electrons, it can be used on samples that are either solid or liquid, conducting or nonconducting. XRF has been widely used to determine wear metals remaining in the oil, as well as separated from the oil. The best sensitivities and widest range of elements are achieved when analyzing the wear metals removed from the oil.

SEPARATION METHODS: There are various well known methods for separating particles from oil. Magnetic separation is commonly used for wear particles, due to the heavy reliance on ferrous structural materials. However, not all components are magnetic, nor are many wear-related contaminants. In some military systems, magnetic materials are eliminated entirely. Centrifugation readily separates metals from oil, but is labor intensive. Electrophoresis and dielectrophoresis form the basis of a number of particle manipulation and separation techniques. Acid dissolution of the suspended particulates has also been used, with XRF analysis performed on the subsequently dried salts.[8] XRF has also been carried out on pellets prepared from thermal decomposition products prepared without regard to settling, evidently on samples containing only very fine particles.[9]

Generally, the most convenient and commonly used method of separation is filtration. Thin planar filters, called membrane filters, are suitable for separation of wear particulates from drawn oil samples of small volume and have been so used by many workers engaged in wear particle analysis. However, they have a limited loading capacity. A semi-automated, quantitative filtration apparatus has been developed for use with drawn oil samples and is discussed in the succeeding section. A specialized diagnostic filter for particle studies has been offered by Pall, from which a layer of filter medium may be removed manually to recover the particulates.[10]

Depth filters collect the particulates within a labyrinth of filter media and are much more suitable for in-line use, although the particles must then be recovered from the filter media prior to performing wear particle analysis. The Canadian Air Force and the GasTOPS diagnostic instrumentation firm are nearing completion of the development of an automated

aircraft filter debris recovery system known as the Deployable Filter Debris Analyzer (DFDA). Debris is dislodged from the filter through a back-flushing process of alternating oil and air streams in ~15 minutes. Debris ≥100 μm in size is quantified for size, number, and type by an included GasTOPS MetalSCAN sensor, and the data are stored or downloaded. Conductive particles traveling through the sensor are identified as ferromagnetic or nonferromagnetic. Ferromagnetic debris is magnetically extracted from the stream, while the remaining debris is deposited on 200 and 20 micron mesh patches for further analysis off line as desired. The ferromagnetic debris is later released onto a second set of patches.

XRF FILTER DEBRIS ANALYSIS (FDA): XRF has been performed in conjunction with FDA much less frequently than it has with particles suspended in oil, even though it offers increased sensitivity as well applicability to all the structural metals with atomic numbers as low as magnesium. FDA allows the preconcentration of the debris from any drawn oil volume. Removal from the oil eliminates any contribution from possible metals in the lubricant additive package, as well as undesirable effects in the x-ray measurement having to do with x-ray absorption and x-ray scatter by the oil.

The value of filter debris analysis (FDA) by XRF was clearly established as early as 1977 by C.A. Waggoner, H.P. Dominique, and K.I. McRae, who sought to understand why a CH-124 Sea King helicopter being monitored by AAS with sampling interval of ~15 hrs. had its turbine engine bearing fail when the AAS results had indicated subcritical wear for Fe and Cu. (Such failures of optical spectroscopic methods are also reported elsewhere.[11]) They re-analyzed the oil samples post facto by AAS and found the same results. They then compared with other methods that measured total metals present in the oil. Total wear particulates were determined by direct filtration of 1 milliliter aliquots through 1.2 micron filters, followed by residue dissolution of the filter debris in nitric acid and AAS analysis of the solution for Fe, Cu, and Ag. The filter debris from 1-2 milliliter samples were also analyzed for total wear particulates by XRF prior to acid digestion, which gave results in agreement with the acid digestion analysis. These total methods were applied to the same samples as had originally failed to provide warning. Both the XRF FDA and the *acid digestion AAS* total methods showed that satisfactory warning of the impending failure would have been obtained, had either of these procedures been carried out instead of AAS performed in the usual direct manner. The direct AAS results were approximately half the values of the total methods in this case. At issue is the inability of *direct AAS* to measure the larger particles characteristic of accelerated wear, which is rectified by the FDA methods applied by these authors. XRF FDA is simpler than acid digestion methods and proved just as accurate as any of the total methods used, while surpassing the optical spectroscopic method.

While the work of Waggoner, et al., is important because of its implications for maintenance management, the report by H. Meier, et al., is distinctive for its breadth of XRF excitation and detection technology applied to wear particles.[12] Meier, et al., drew samples of several milliliters from diesel automotive and truck engines and analyzed them by XRF in liquid and filtered forms. They performed trending of wear, most commonly but not exclusively for iron and copper. Their most remarkable results were obtained from a bearing test rig on which they monitored bearing load and wear rate. The wear rates were determined by XRF FDA.

The graphed results[13] show typical Fe wear rates of several 10s of μg/h, with an increasing trend following the appearance of a large particle. Removal of the load dropped the wear rate nearly 100x to below 0.5 μg/h, which recovered to over 100 μg/h upon reinstating the load.

The wear rate results of Meier, et al., demonstrate the high sensitivity and wide dynamic range of XRF FDA, even when applied to a few milliliter of drawn sample. They also show the ability of XRF FDA to detect changes in operating conditions of the bearing, and the acceleration of wear accompanying large particle release by a bearing.

M. K. Greenberg and F. M. Newman were very concerned about two sample preparation factors: (1) x-ray self-absorption due to particle size, and (2) x-ray self-absorption due to particle layer thickness.[14] They listed a third factor, the uniformity of the deposit, but this just restates the second factor in terms of local thickness. They found, as expected, that controlling these factors, with a deposit of under 2 mg for typical x-ray system geometries and sample sizes, reduces interelement matrix effects and particle size effects to negligible levels. Their approach to controlling the particle size, however, was rather tedious and involved hand grinding the debris, recovered from drawn samples, to under 2 μm sizes. No data on actual machine wear was presented, although the authors claim to have applied the method extensively to turbine engines.[15]

A man-portable, semi-automatic Oil Wearmetal Analysis by X-ray (OWAX) system, employing XRF filter debris analysis of drawn oil samples, has been developed for the US Air Force.[16], [17], [18], [19] The drawn oil sample of 3 milliliter is introduced into the apparatus, and filtered through a 0.4 micron pore membrane filter by a quantitative filtration apparatus (QFA) that also rinses the particles. Upon manual action, the membrane-filtered particulates are mechanically presented by the QFA to the XRF unit. Five structural metals are fully analyzed by XRF simultaneously in less than four minutes per sample, with filtration complete in the first 100 seconds. For high concentrations of wear metal, the unit is capable of direct XRF analysis of suspended wear metal in a 0.3 milliliter oil sample.[17] Proposed enhanced capabilities include a vacuum path to enable analysis for Al, Si, and Mg.

A recent study by G. Humphrey of the US military Joint Oil Analysis Program Technical Support Center (JOAP-TSC) advanced the field considerably by relating XRF FDA analyses of a population of F404 turbine aircraft engines to the engine metallurgy and the entries in the maintenance data base for those engines.[20] Individual engines and engine modules were tracked and compared with the XRF analysis. No new hardware was developed and all apparatus are commercially available. Two US Navy airfields removed oil filter cartridges as they normally became available (expected filter life is 200 hours) and sent 189 of them to the JOAP-TSC. At the JOAP-TSC, each cartridge was placed in a solvent and sonicated for a set time, and the released debris were then re-filtered onto a membrane for insertion into a TN Spectrace automated XRF system. The XRF instrument measures, computes, and reports the quantity of each of approximately two dozen elements present in the sample. The entire procedure takes ~20 minutes, or ~15 minutes per cartridge when several are analyzed in succession.

Humphrey categorized the engine condition as reported by the mechanics in the maintenance data base as: normal, foreign object damage (FOD), high time, or failure. Statistical analysis of the body of XRF results for the three most indicative major metals (Fe, Ni, and Ti) produced five levels of statistical significance, with level 5 representing the highest wear metal quantities and level 1 the lowest. It was found that every one of the engines possessing levels 3, 4, or 5 fell into one of the damaged engine categories (FOD, high time, or failure), either at the time the present filter was taken or before the replacement filter's life was exhausted. If XRF indicated the engine was faulty, it was in fact faulty.

Twelve filters were removed from engines that later failed. From these filters, level 3 values or higher (in one or more of the three major metals) were recorded in one engine at 142 hours in advance of failure, at 135 h in another filter, and at 181 h prior to failure in another. Presumably, level 2 may potentially correspond to even greater warning. Level 2 seldom occurred in engines identified by the mechanic as normal, which points to asymptomatic wear in those engines.

Level 2 may be taken as a flag that signals the onset of the final phase of turbine life. While the duration and profile of that final phase is not yet well determined, it is clear that it is significantly greater than 100 hours and that the technical means to monitor the situation can be implemented with XRF. The availability of such an extensive forewarning is an unparalleled opportunity to develop tools to manage the final phase of engine life. Both operations and maintenance could benefit greatly from advances made in this area. With this amount of warning, there is time to make further measurements (perhaps some with other diagnostics), time to plan maintenance outages, and, in general, time to prevent the oncoming failure from becoming a disruption to operations and safety.

Perhaps the most impressive aspect of Humphrey's study is that it was carried out on an engine for which the common diagnostic methods for wear metal analysis have been ineffective. The explanation for this ineffectiveness is simple: the other methods (compositional and otherwise) rely on drawn oil samples that contain no detectible wear metals, because the wear debris are removed from the oil by the fine filtration implemented on these engines. Fine filtration has been found to significantly extend engine life, but heretofore has had the unfortunate side effect of disabling the very diagnostics which give warning when the end of engine life has been reached. XRF FDA offers a hope for a solution to this dilemma by offering the ability to give extensive forewarning before the end of engine life, while retaining the financial advantages of extended engine life.[1]

Early warning of wear is not the only important aspect of Humphrey's JOAP study. By analyzing which elements appear in combination, it is possible to identify which of the engine modules is a likely source of the wear particulates. The metallurgical basis of this analysis was assembled by Humphrey and applied to the XRF results for the F404 engine modules. In the case of the pump oil bearing, Humphrey could carry out the identification to the component level. The metallurgical analysis also figured in his identification of corrosion products, some of them not normally otherwise detected. Likewise, he identified contamination by dust and sand through measurement of the element silicon which, if found at contaminant levels,

dictates remedial action and the possible onset of premature wear.

SUMMARY: Clearly, the wear rate and compositional measurement capabilities offered by X-Ray Fluorescence Filter Debris Analysis match well with the current emphasis on maintenance based on the condition of the specific machine, rather than based on schedules derived from averages or estimations for that type or class of machine. Only by addressing the individual machine condition will early failures be recognized, and ongoing normal life be permitted to continue uninterrupted beyond the scheduled averages. These factors can have a significant impact on the management of operations and maintenance.

REFERENCES:

1. R. R. Whitlock, "Restoring Wear Condition Monitoring with XRF," Proceedings of the 8th Annual Predictive Maintenance Technology National Conference, SC Publishing Division, Minden, Nevada, 26-28, 1996.

2. Peter B. Senholzi, "Oil analysis/wear particle analysis," *Proceedings of the 26th Meeting of the Mechanical Failures Prevention Group*, T. Robert Shives and William A. Willard, eds., National Bureau of Standards, Washington, D.C., 27-32, 1977.

3. J. A. Alcorta, J. H. Mohn, and L. L. Packer, "Improved capabilities to detect incipient bearing failures," AFWAL-TR-80-2057, July 1980.

4. Piran Sioshansi, "Surface layer activation technique for monitoring and in-situ wear measurement of lubricated systems," *JOAP International Symposium Proceedings*, Joint Oil Analysis Program Technical Support Center, Pensacola, Florida, TR-83-12, 103-117, 1983.

5. P. Sioshansi and F. L. Milder, "Computer Based Wear Analyzer for Surface Layer Activation Measurements," *Proceedings of the 36th Meeting of the Mechanical Failures Prevention Group,* Whittaker, G. Allan with T. Robert Shives and Gerald J. Philips, eds., 118-26, 1982.

6. Graham F. Fisher and L. G. I. Bennett, "Quantitative Filter Debris Analysis (QFDA) As a Means of Monitoring Wear in Oil-Lubricated Systems," Journal of Radioanalytical and Nuclear Chemistry **180**/1, 121-129, 1994.

7. See R. R. Whitlock, to be published in forthcoming volume in the series, *Advances in X-Ray Analysis.*

8. C. K. Bhat, D. Kaul, G. S. Lodha, R. Koul, and K. J. S. Sawhney, "Wear and tear studies of oil-lubricated machines using energy-dispersive x-ray fluorescence spectrometry," X-Ray Spectrometry, **18**, 243-5, 1989.

9. Yu-Wen Liu, A. R. Harding, and D. E. Leyden, "Determination of wear metals in oil by x-ray spectrometry," *Advances in X-Ray Analysis*, Vol. 29, Plenum Press, New York, 503-7, 1986.

10. P. V. Madhavan, "Monitoring aircraft power plant and transmission systems via diagnostic filters," *Proceedings of the 43rd Meeting of the Mechanical Failures Prevention Group*, 44-47, 1991.

11. W. E. Rhine, C. S. Saba, and R. E. Kauffman, "Metal particle detection capabilities of rotating-disk emission spectrometers," Journal of the American Society of Lubrication Engineers **42**/12, 755-761, 1985.

12. H. Meier, E. Unger, W. Albrecht, N. Geheeb, W. Hecker, U. Tschirwitz, and D. Bösche, "Untersuchungen zur Schadensfrüherkennung bei Dieselmotoren mit Hilfe der energiedispersiven Radionuklid-Röntgenfluoreszenzanalyse," Bundesministerium der Verteidigung, Forschungsbericht aus der Wehrtechnik BMVg-FBWT-79-19, Staatliches Forschungsinstitut für Geochemie, Bamberg, Germany, 1979.

13. H. Meier, op. cit., Fig. 54.

14. M. K. Greenberg and F. M. Newman, "Application of Energy Dispersive X-Ray Fluorescence Spectroscopy to the Analysis of Contaminants In Fuels and Lubricants," AFLRL No. 102, Nov. 1978.

15. Although Greenberg and Newman claim to have applied this methodology to over 2000 turbine engine samples for the US Army, no documentation of that work could be found by a librarian.

16. B.C. Clark and B.J. Cook, "Oil Wearmetal Analysis By X-Ray (OWAX)," AFWAL-TR-82-2020; MRC-82-509, May, 1982.

17. Benton C. Clark, Sidney L. Russak, Michael G. Thornton, and B. Judith Cook, "An instrument for oil wear metal analysis by x-rays (OWAX)," in *JOAP International Symposium Proceedings*, Joint Oil Analysis Program Technical Support Center, Pensacola Forida, 162-173, 1983.

18. Benton C. Clark, Vincent P. Woerdeman, Michael G. Thornton, B. Judith Cook, Phillip W. Centers, and Warren C. Kelliher, "A Portable X-Ray Analyzer For Wearmetal Particles in Lubricants," in *Proceedings of the 36th Meeting of the Mechanical Failures Prevention Group*, Whittaker, G. Allan with T. Robert Shives and Gerald J. Philips, eds., Cambridge University Press, Cambridge, 189-197, 1983.

19. Michael G. Thornton, "Portable x-ray fluorescence spectrometer for environmental monitoring of inorganic pollutants," US Patent 5,014,287, 7 May 1991.

20. G. R. Humphrey, "Characterization of Debris from F404 Engine Oil Filters by Energy Dispersive X-Ray Fluorescence," JOAP-TSC-TR-96-02, Joint Oil Analysis Program Technical Support Center, Pensacola, Florida, 14 June 1996.

CONDITION MONITORING OF PERFLUOROPOLYETHER (PFPE) LUBRICATED COMPONENTS

Regan L. Silvestri, Ph.D.
Senior Research Scientist

PREDICT/DLI
PREDICT Technologies Division
9555 Rockside Road
Cleveland, Ohio 44125

ABSTRACT: *Perfluoropolyethers (PFPEs) are highly resistant fluids used as lubricants in demanding applications. Condition monitoring of PFPE lubricated components is fundamentally different from condition monitoring of components lubricated with traditional hydrocarbon oils. Analogous to more conventional condition monitoring, the condition of both the lubricant and the machine can be monitored via used oil analysis and wear particle analysis, respectively. However, modifications on industry standard laboratory techniques are necessary to analyze these perfluorinated fluids. Moreover, the laboratory data must be interpreted with an understanding of these distinct systems. This paper addresses the nuances of condition monitoring of perfluoropolyether lubricated components.*

Key Words: Condition monitoring, Ferrography, Fourier Transform-Infrared Spectroscopy, Perfluoropolyether, Used oil analysis, Vacuum pump, Wear particle analysis

INTRODUCTION

Due to the strong nature of the carbon-fluorine bond, perfluoropolyethers (PFPEs) are highly chemically resistant lubricants. They are therefore used in demanding applications where chemically resistant lubricants are needed due to the possibility of attach by harsh process contaminants. In addition, PFPEs are used in specialty high value applications. For example, PFPEs are used as vacuum pump fluids in semiconductor production on account of the low level of backstreaming that they provide. Condition monitoring of components lubricated with perfluoropolyethers is essential due to the high value applications that use these costly lubricants.

Analogous to conventional condition monitoring, the condition of the lubricant can be monitored via used oil analysis and the condition of the machine can be monitored via wear particle analysis. Useful wear particle analysis techniques include Direct

Reading Ferrography and Analytical Ferrography. The lubricant condition can be monitored via used oil analysis techniques including viscometry, FT-IR spectroscopy, trace elemental analysis and the crackle test. However, modifications on these industry standard laboratory techniques are necessary to analyze PFPE fluids. As the most basic example, fluorinated solvents must be used to dissolve the perfluorinated lubricants. In addition to the modified laboratory techniques, the laboratory data must be interpreted with an understanding of these distinct systems. That is, the used oil analysis data must be interpreted with consideration to the fluorine chemistry involved. Additionally, the wear particle analysis data must be interpreted with an understanding of these special lubricated components, their wear modes and metallurgy.

VISCOSITY

The viscosity of a lubricant is probably without question the single most important property of a lubricant. The viscosity of the oil must be correct for the component's operating load, speed and temperature.

When testing perfluoropolyethers standard glass capillary viscometers should be used to determine viscosity. Kinematic viscosity is thus determined by measuring the length of time for the sample to flow through the glass capillary. The glass capillary tubes are immersed in constant temperature baths to control the temperature at which the viscosity is measured. The length of time for the sample to flow through the tube is directly proportional to the kinematic viscosity through the calibration constant by the equation

$$\nu = C\,t$$

where the calibration constant is unique for each capillary tube.

Our laboratory uses Zeitfuchs cross-arm tubes. However, any tube geometry that is designed for use with clear fluids would be suitable. The difference in running PFPE oils as compared to hydrocarbon oils is in the cleaning of the capillary tubes between samples. When the testing has been completed for a sample a fluorinated solvent must be used to clean and rinse the tube as the PFPE sample will not dissolve in standard hydrocarbon solvents. There are currently many suitable perfluoro solvents and hydrofluoro solvents on the market as replacements for chlorofluorocarbons. Several perfluoro solvents and hydrofluoro solvents are available from DuPont, 3M and Ausimont; and new "Freon alternates" are being introduced persistently. We have evaluated many such solvents and find several to be suitable.

After cleaning the capillary tube with a fluorinated solvent a volatile hydrocarbon solvent such as Acetone is used to dry the tube before introducing the next sample. This avoids a diluting effect of residual solvent on the next sample.

It should be mentioned that methods other than glass capillary viscometry may not be suitable for PFPE samples. Many such other methods for measuring viscosity are accurate only for Newtonian fluids and PFPEs exhibit non-Newtonian behavior.

Furthermore, many such other methods for measuring viscosity measure dynamic viscosity and not kinematic viscosity. The two are related through density so results can be converted if the density is known. However it must be accounted for that the density of PFPEs are in the range 1.9g/mL whereas the density of mineral oils are typically 0.87g/mL, and many of the instrument for measuring viscosity and converting from dynamic to kinematic viscosity are "hard coded" with a density of 0.87g/mL because they are designed for petroleum based oils.

CRACKLE TEST

The crackle test is a simple test where a sample of the lubricant is placed directly onto a laboratory hot plate to test for water and volatile contaminants. A sample of approximately 5mL is dropped onto a hot plate that is maintained at a temperature just above the boiling point of water. The temperature of the hot plate is usually maintained at about 120°C as measured with a surface thermometer. At this temperature any water in the oil will boil and the sample will appear to "pop." The water will pop in only a few localized areas, but the pops will have a large radius. Any volatile gasses will also evaporate but their appearance is very different from that of water. Volatile gases evaporate with more of a "crackling" sound. The crackling is not localized but more widespread and the radius of the bubbles is smaller.

Although it has the outward appearance of being a crude and "un-technical" laboratory test, we have had outstanding success with the crackle test. We have found the crackle test to be very sensitive to water well below concentrations of 1000ppm. This test is, of course, subjective and therefore relies on the experience of the operator.

FOURIER TRANSFORM-INFRARED SPECTROSCOPY

Fourier Transform-Infrared (FT-IR) Spectroscopy has been successfully used to detect PFPE oxidation, water contamination, and contamination with particulates. The FT-IR spectrum must be collected in transmission as reflection techniques do not have the sensitivity needed to detect these low levels of degradation and contamination. We collect spectra in transmission using a 100μ sealed liquid cell with ZnSe windows. We co-add 16 interferograms and use a spectral resolution of $8cm^{-1}$, however both of these parameters can be varied.

It is important to realize that the chemistry of these perfluorinated samples is very different from that of petroleum based lubricants. Therefore, the FT-IR spectra of PFPEs is very different from that of mineral oils. Furthermore, the degradation/oxidation mechanism of PFPEs is very different from that of mineral oils. Whereas mineral oils oxidize to form a variety of carbonyl species, PFPEs oxidize differently. This is extremely significant because the FT-IR spectrum must be analyzed in the proper spectral regions to detect oxidation products that are unique to perfluoropolyether lubricants.

Figure 1 shows an example of several overlaid FT-IR spectra in the region of the spectrum that is indicative of lubricant oxidation. Note in Figure 1 that two samples show various levels of oxidation while the other four samples do not show oxidation.

Figure 2 shows several overlaid FT-IR spectra in the region of the spectrum that is indicative of water contamination. Note the peak is the spectrum of the one sample indicating water contamination that is not in the spectra of the other samples. Again, the area of the peak indicates the relative level of water contamination.

The final data point provided by FT-IR spectroscopy indicates the level of total particulate contamination in the fluid. Any particulates in the lubricant act to diffract the infrared light causing an upward shift and a sloping of the baseline in the IR spectrum.

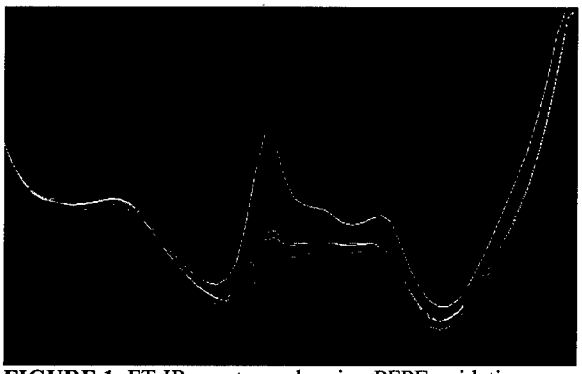


FIGURE 1: FT-IR spectrum showing PFPE oxidation.

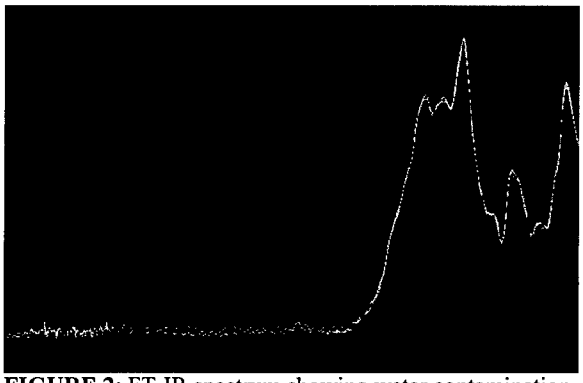


FIGURE 2: FT-IR spectrum showing water contamination.

As one example of particulate contamination it has been our experience that a failing seal causes the lubricant to become cloudy. Whether this clouding of the fluid is due to oxidation of the lubricant or due to the process dissolved in the lubricant is not known. But in either case, the clouding does indicate a failing seal and can thus be used to diagnosis this condition.

FT-IR results are reported by integrating the area of the peak to generate a value in units of Infrared Absorption (Abs). Since the units of Abs are relative and not absolute, sample values are most informative when compared to baseline values for a new un-used lubricant. There is information in a single Abs value for a sample, however the information is more valuable when this relative number is (a) compared to the normal value for the same un-used lubricant, and (b) trended through time for a particular sample point.

TRACE ELEMENTAL ANALYSIS

Emission Spectroscopy is used to quantify the concentrations of 21 different elements including wear metals and contaminants. Traditionally this technique is also used to monitor additive levels, however PFPEs are neat fluids. The most common form of emission spectroscopy in condition monitoring is Rotating Disk Electrode (RDE) spectroscopy. A porous carbon disk rotates through a sample of oil where the various contaminants and wear particles are deposited onto the disk. The disk continues to rotate and the sample is pulled up and into an electrode gap where a high energy charge vaporizes or "burns" the sample. As the various elements "burn" they emit light at characteristic frequencies. The intensity of emitted light is measured at each frequency of interest and calibrated to elemental concentrations in units of ppm. It is important to realize that the results displayed in units of ppm are not absolute concentrations but relative concentrations due to incomplete vaporization of the solid particles. However, this does not lessen the value of this technique in condition monitoring as accurate historical trends can be developed and deviations from these trends represent significant conditions.

The laboratory procedure for running PFPE samples on the RDE spectrometer is very similar to that for running traditional mineral oil samples. However, there are some important differences in calibrating the spectrometer, and many important differences in the software program that operates the spectrometer. We have set up two programs on the computer that operates our RDE spectrometer so that the laboratory personnel may simply switch back and forth between the two programs. The program for analyzing PFPE samples carries with it not only the parameters for running the analysis but also the calibration information that was run specifically for these samples.

The differences in the program to analyze PFPE samples are again centered around the fact that these fluids are perfluorinated. The first difference is due to the fact that the fluorine lines are very strong due to their high concentration. These intense fluorine lines may interfere in areas of the spectrum where other elements are normally detected. Therefore, lines for the various elements must be chosen that are free of interferences from fluorine. This is easily accomplished as each element emits light at a variety of frequencies so one can choose a suitably intense line for an element that is free of interferences from fluorine and other elements. Furthermore, an element is often set up to measure at two different frequencies, one frequency at lower concentrations and another frequency at higher concentrations. The program must therefore be set up to switch between two lines for an element that are both free of interferences.

The second difference in analyzing PFPE samples is centered around the fact that PFPEs have a different background emission as compared to hydrocarbon oils. This different background emission results in a spectrum baseline that is different from that of mineral oils. This is dealt with by two modifications to the standard analysis for hydrocarbon oils. First, baseline corrections are set up appropriately. That is, each peak intensity is measured as the intensity of the peak above the spectrum baseline. To set up this baseline correction appropriately each peak is observed and a corresponding baseline area is located where the baseline intensity can be measured for that peak. Once this baseline area is located and programmed into the method there is no need to observe the individual peaks in the routine sample analysis. Therefore, analysis of a PFPE sample is no more troublesome than the analysis of a mineral oil sample. The second modification in running PFPE samples that is due to their different background emission manifests itself in calibrating the instrument with a 0ppm standard. The traditional 0ppm standard of mineral oil has a background emission very different from that of a pure 0ppm standard of PFPE. Therefore, if the spectrometer were calibrated with the traditional 0ppm standard of mineral oil and PFPE samples were tested, the samples would be tested against an incorrect background. The solution is simply to calibrate the 0ppm level with a standard of pure PFPE since this will have the same background emission as the PFPE samples to be tested.

The third and final difference in analyzing PFPE samples is based upon the fact that PFPEs, being perfluorinated, do not contain any hydrogen atoms. When a sample in analyzed and the measured light intensity is converted to a concentration, that concentration is reported in units of ppm. This implies that if a certain amount of an element is detected, the total amount of sample must be known so that the concentration of the element can be reported as a fraction of the total sample. With mineral oil samples when a sample is analyzed the intensity of the carbon and hydrogen lines are also measured. Therefore, the ratio of the intensity of the wear metal peak relative to the hydrogen or carbon reference allows reporting of the concentration of the wear metal in units of ppm. Since PFPE samples do not contain any hydrogen all elements must be referenced to carbon. Once this modification is made to the program for analyzing PFPE samples the analysis of a PFPE sample is no more troublesome than that of a mineral oil sample.

DIRECT READING FERROGRAPHY

Above has been described techniques that largely reveal the condition of the lubricant. Perhaps more important than the condition of the lubricant is the condition of the equipment or machine. The condition of the equipment is monitored by analyzing the machine wear particles, and this is perhaps more important than monitoring the condition of the lubricant as most failures are traceable to equipment or machine conditions. Useful wear particle analysis techniques include Direct Reading Ferrography and Analytical Ferrography.

Direct Reading Ferrography is a technique for trending relative amounts of ferrous machine wear and detecting increases that indicate abnormal machine wear

modes. A magnet is used to collect ferrous wear particles as an oil sample flows through a disposable glass capillary tube. Light is shined through the capillary tube and the amount of light blocked is measured on the opposite side of the capillary tube. The intensity of light attenuated is directly related to the amount of wear debris deposited in the path of the light. As such, the relative value for light intensity that is measured can be trended through time and increases in the amount of machine wear can be detected.

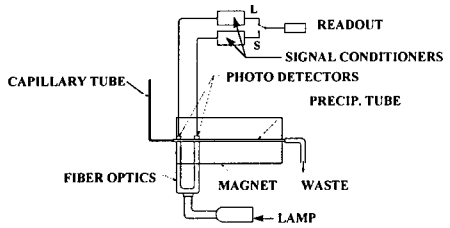


FIGURE 3: Schematic of the Direct Reading Ferrograph.

The capillary tube is placed in the magnetic field such that the oil sample flows through the magnetic field. Larger wear particles are trapped as soon as they enter the magnetic field whereas smaller wear particles may flow further into the capillary tube. By measuring at two points in the magnetic field the density of large particles (D_L) and the density of small particles (D_S) can be distinguished. This is an important distinction because it is the large particles that typically indicate critical or abnormal wear modes. The D_L and D_S are summed to calculate the total Wear Particle Concentration (WPC), the key parameter that is trended over time.

Components lubricated with perfluoropolyethers commonly generate small amounts of machine wear. Therefore, any technique used to monitor machine wear must be very sensitive and Direct Reading Ferrography accomplishes this. Vacuum pumps, for example, usually operate with Wear Particle Concentrations (WPCs) below 5.

One again, the procedure for analyzing PFPE samples in the laboratory is much like analyzing hydrocarbon samples. The only difference is that fluorinated solvents must be used to dissolve a PFPE sample.

Figure 4 shows a typical example of how the Wear Particle Concentration is trended over time. Note the increasing trend and how the fourth sample has exceed the statistical alarm limit. The statistical alarm limits are calculated automatically and updated with each new historical sample by software. These alarm limits are used to detect statistically significant changes in wear particle trends. DR Ferrography is an extremely valuable technique in detecting abnormal wear modes; however, it does not diagnose the cause of the critical wear mode. For this we must turn to more advanced analysis by Analytical Ferrography.

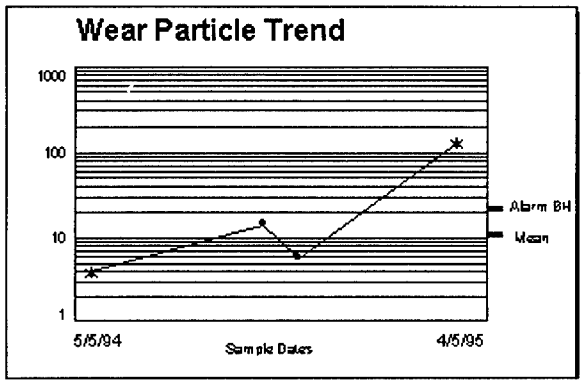


FIGURE 4: Example of a Wear Particle Concentration trend by Direct Reading Ferrography for a Vacuum pump.

ANALYTICAL FERROGRAPHY

Once an abnormal wear mode has been detected the condition must be diagnosed so that the proper maintenance action can be recommended. Analytical Ferrography is a technique whereby the machine wear particles are deposited on a glass microscope slide and then viewed with an optical microscope. Figure 5 shows a schematic of a Ferrogram. A new Ferrogram is placed at a specified incline in a magnetic field and a diluted oil sample is pumped down the slide at a designated flow rate. The magnet acts to trap the ferrous wear particles is an orderly and systematic fashion. The non-ferrous particles are deposited on the Ferrogram as they become lodged by the ferrous particles due to the low flow rate. Once the sample is completely pumped over the Ferrogram solvent is pumped down the Ferrogram to rinse away residual oil so that when the solvent dries the Ferrogram is clean of oil with the wear debris fixed to the slide. This slide is then analyzed under an optical microscope to determine the size, shape, concentration and composition of the wear particles.

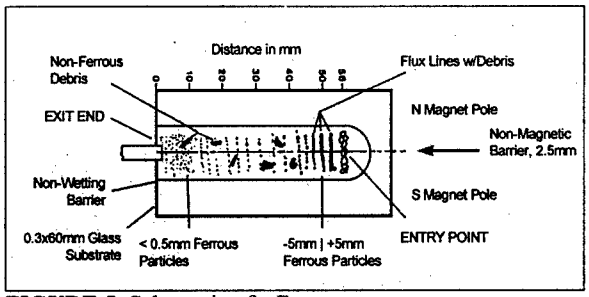


FIGURE 5: Schematic of a Ferrogram.

The size of the wear particle is important because it is the large particles that typically indicate critical wear modes. The shape of the wear particle diagnoses the wear mode that created the particle thus indicating what maintenance action should be taken. The concentration or amount of a particular type of wear particle indicates how severe the wear mode has progressed. Finally, the composition or metallurgy of the wear particle indicates the wearing component which locates where in the equipment maintenance should be performed.

Preparation of Analytical Ferrograms with PFPE samples proceeds in the laboratory much like the preparation of Ferrograms for hydrocarbon samples. However, once again a PFPE sample must be diluted with a fluorinated solvent. Also, the slide must be rinsed with a volatile fluorinated solvent to rinse away the perfluorinated sample. Finally, a larger sample size is usually used to deposit a significant quantity of wear debris. We have set guidelines for determining the amount of sample that is used. Despite the fact that an increased sample size is usually used for Analytical Ferrography, Ferrograms still contain small amounts of wear. This is simply an indication of the high level of engineering in the design of vacuum pumps and other similar machines.

Figure 6 shows some examples of wear particles that have been observed on Analytical Ferrograms of vacuum pumps. In general, the same modes of wear are observed in these components as other similar components. Figure 6A shows 5μ low alloy steel normal rubbing wear. Note the small amounts of this normal and benign wear that are generated. Figure 6b shows a 20μ stainless steel bearing wear particle. This type of bearing particle indicates bearing fatigue spalling. Figure 6c shows a large 100μ copper sliding wear particle. Sliding wear indicates excessive loads or speeds at critical contact points. One unique type of material that is observed on Ferrograms of PFPE lubricated components is an amorphous gel that is believed to be oxidized lubricant.

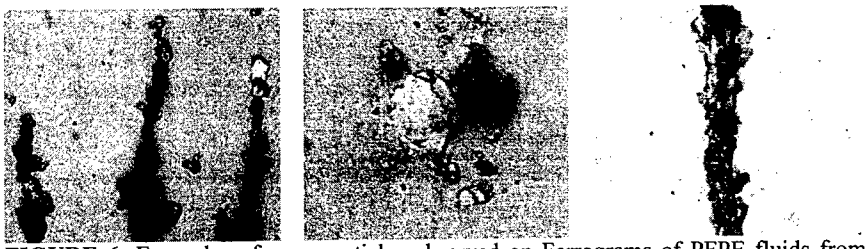


FIGURE 6: Examples of wear particles observed on Ferrograms of PFPE fluids from vacuum pumps; (6A) normal rubbing wear, (6B) bearing wear, (6C) sliding wear.

TESTS NOT PERFORMED

The Total Acid Number (TAN) and Karl Fischer (KF) tests, commonly performed on hydrocarbon oils, are not recommended for PFPE lubricants. The TAN is a test for lubricant acidity that increases as an indication of lubricant oxidation. KF is a test that quantifies the water content of the lubricant. Although they are relevant properties, these

tests are difficult to perform on PFPE samples because both tests are based on laboratory techniques that measure the lubricant sample's electrical conductivity, which is troublesome for these dielectric samples.

It is not possible to perform the Karl Fischer test on PFPE samples for two reasons. Firstly, commercial KF reagents are formulated in hydrocarbon solvents. These solvents do not dissolve the PFPE sample. Secondly, the PFPE sample is di-electric and thus does not conduct electricity. Karl Fischer titration endpoints are detected coulometrically or potentiometrically, and this type of electrical end point detection is not possible when the sample does not conduct. It is not possible to perform the TAN test on PFPE samples for the same two reasons; the sample does not dissolve, and a potentiometric endpoint can not be detected.

The fact that these two standard tests can not be performed is not of concern. The KF test is a test for water contamination and this information is furnished in the FT-IR test and the crackle test. The TAN test is a test for lubricant oxidation and this information is more directly furnished from the FT-IR test as the TAN test is an indirect test for oxidation because it senses the oxidation byproducts. Therefore, both the KF test and the TAN test are replaced by the single FT-IR test.

SUMMARY

There are many possible testing methods for condition monitoring. It is not appropriate to perform all of the possible tests but only the relevant tests. With the experience to gain the knowledge of what failure modes these types of equipment suffer, condition monitoring tests can be performed that test for these relevant failure modes. We have therefore designed a set of tests specific for monitoring PFPE lubricated vacuum pumps that we refer to as an Equipment Specific Analysis. This Equipment Specific Analysis includes: viscometry, FT-IR spectroscopy, the crackle test, trace elemental analysis, Direct Reading Ferrography, and Analytical Ferrography. Viscometry, FT-IR spectroscopy, the crackle test, and trace elemental analysis are used to monitor the lubricant's condition. Direct Reading Ferrography, Analytical Ferrography, and trace elemental analysis are used to monitor the machine or component's condition. Furthermore, it has been our experience that equipment condition monitoring is more valuable than lubricant condition monitoring as most failures result from equipment conditions and not lubricant conditions.

Moreover, we have gained considerable experience with these types of equipment, most notably vacuum pumps, and have learned that most dry pump failures can be traced back to a root cause of seal failure and most wet pump failures can be traced back to a root cause of corrosive process attack on the lubricant. In a dry pump the failing seal exposes the harsh process chemistry to the lubricant resulting in oxidation of the lubricant. In a wet pump where the process is not separated from the lubricant exposure of the harsh process to the lubricant oxidizes the lubricant. In both cases oxidation of the lubricant creates corrosive acids that attack the interior metal surfaces of the pump. Also, oxidation of the lubricant increases the viscosity of the lubricant which results in

increased machine wear. This makes the FT-IR test a very important part of this testing package as it is the FT-IR test that trends oxidation of the lubricant.

In conclusion, the FT-IR test is a valuable condition monitoring tool in the analysis of PFPE vacuum pump fluids because it is sensitive to lubricant oxidation, which is the root cause of a majority of vacuum pump failures. Furthermore, this single laboratory test has the advantage of providing three independent data points indicative of (a) lubricant oxidation, (b) particulate contamination, and (c) water contamination.

ROTRODE FILTER SPECTROSCOPY REVISITED - A METHOD FOR MULTI-ELEMENTAL ANALYSIS OF PARTICLES IN USED OIL

Daniel P. Anderson and Malte X. Lukas
Spectro Incorporated
160 Ayer Road
Littleton. MA 01460

Abstract: A convenient and economical method has been developed to provide spectrometric analysis of the wear and contaminant particles in used oil samples. It has been recognized for some time that spectroscopy is more or less blind to the larger particles in an oil sample, precisely those particles which are sometimes most indicative of an abnormal wear mode. The rotrode filter spectroscopy (RFS) method makes use of the fact that the carbon disc electrode used in a rotating disc electrode (RDE) spectrometer is itself porous and may be used as a filter. A fixture was designed so that used oil sample is pulled through the disc which is then washed with solvent thus retaining only particulates on the outer circumference of the disc. The disc is subsequently "burned" on a spectrometer providing an analysis for larger particulates in a sample as apart from molecular material and very small particles in the oil sample as would be obtained by conventional spectrometric analysis. The RFS method is gaining acceptance in commercial and industrial laboratories. The method and an example of its application in a predictive maintenance program will be described.

Key words: Ferrography; oil analysis; rotating disc electrode (RDE) spectrometers; rotrode filter spectroscopy (RFS); wear particles

Introduction: Spectrometric oil analysis has been applied for more than 30 years as a routine and cost-effective condition monitoring technique. It is used to determine the elemental concentration in parts per million of wear metals, contaminants and additives in an oil sample. Commercial oil analysis laboratories report on as many as 20 different elements. With prior knowledge of the wear metal and contaminant limits for the machine or engine being evaluated, it may be determined if that equipment is operating properly or if preventive maintenance is required.

A fact which has recently become more widely known is that spectrometric oil analysis is not responsive to large particles in an oil sample. This is true to a greater degree for certain types of spectrometers than others. It is also more important for certain types of machines.

Atomic absorption (AA) and inductively coupled plasma (ICP) spectrometers suffer the most from particle detection inadequacies. Rotating disc electrode (RDE) spectrometers are responsive to somewhat larger particles, but the upper limit is still approximately 10 μm.

It was known as long ago as the early 1970's that certain types of failure modes were not easily detected by spectrometric oil analysis. Severe wear modes such as spalling, severe sliding wear and cutting wear generate large particles which may go undetected by spectroscopy. In fact, ferrography was developed at that time with financial support from the Office of Naval Research to overcome the detection inadequacies of spectrometric oil analysis in regard to rolling element bearing failures in aircraft gas turbine engines.

Rotrode Filter Spectroscopy (RFS): A fixture, Figure 1, was designed to clamp the carbon electrode discs so that oil is drawn through the outer circumference of the discs when a vacuum is applied to the inside of the discs. [1] The particles in the oil are captured by the disc. The oil is then washed away with solvent and the disc is allowed to dry. The disc is removed from the preparation fixture and transferred to the RDE spectrometer where the particles captured by the disc are vaporized and detected. A multi-station fixture is used so that a number of samples can be filtered at once. The procedure is fast, and therefore economical to perform.

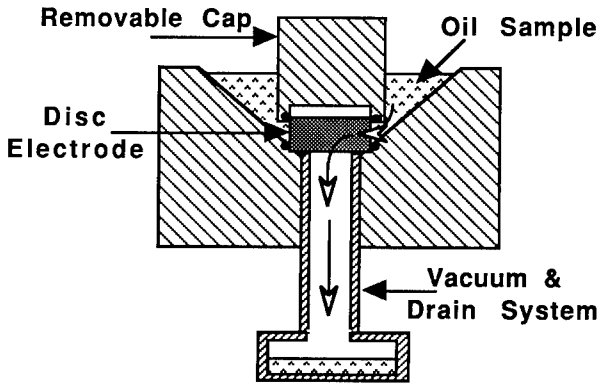


Figure 1, Diagram of RFS Sample Preparation Fixture

During initial development of the RFS method, a test sample was prepared by carefully weighing commercially available metal powders into base oil (0 ppm oil) so that the weight concentration of metal powders in the oil was known. Nickel, chromium and molybdenum powders were used. A number of disc electrodes were prepared by filtering one milliliter of test sample through a disc electrode and then rinsing the oil away by filtering 3 ml of solvent (hexane was used although other solvents may be used). The solvent is applied 1 ml at a time to reduce oil residue in the bowl of the filtration fixture. The disc electrodes were then analyzed on an RDE spectrometer.

The resulting intensities for the various spectral lines were normalized to the intensity of a carbon reference line. Referencing to carbon improves repeatability because it corrects for the gross intensity of individual burns. The intensity ratio of the blank disc is

subtracted from the intensity ratio of the test sample for each spectral line to give the intensity ratio due only to the presence of analyte.

It was found that the signal for the filter discs prepared by filtering 1 ml of test sample with Cr, Ni and Mo particles was greater than the signal for an equivalent concentration of organo-metallic standards run in the usual manner. The amount of signal for a given amount of sample determines the sensitivity of the technique and is related to the limit of detection. Table 1, below, shows data for a test sample with Cr, Ni and Mo particles. The BEC (Background Equivalent Concentration) is the concentration of analyte that would give an intensity (signal) equal to a blank sample. The lower the BEC, the lower the Limit of Detection (LOD).

Table I, Tests with Commercial Powdered Metal Particles in 1 to 5 μm Size Ranges

	Cr	**Ni**	**Mo**
Concentration, ppm	11.4	15	11.2
BEC for Normal Method, ppm	14.15	14.23	10.53
LOD for Normal Method, ppm	0.71	0.71	0.53
BEC for RFS, in ppm	6.00	1.39	5.05
LOD for RFS, in ppm	0.30	0.07	0.25

The sensitivity and limits of detection of the RFS method can be improved by filtering more sample through the disc electrode. It is an advantage of the RFS method that it is an integrative technique. More sample volume can be processed to determine smaller and smaller quantities of analyte. Twice as much sample cuts the limit of detection in half, and so on. Filtering larger sample volumes can be applied quite readily to many types of samples, although diesel engine lubricating oils tend to clog the disc electrodes depending upon soot content. Consequently, there is a practical limit to how much diesel lubricating oil sample can be filtered. From a practical standpoint, it takes 5 or 10 minutes to filter a typical sample, although diesel oil samples can take substantially longer.

During development of the RFS method it was determined that:

1) Results for blank discs were not significantly different from results obtained from discs through which base oil (0 ppm standard) was filtered and washed away by the solvent rinse and

2) Filtration of 100 ppm organo-metallic standard through a disc followed by the solvent rinse causes only a slightly elevated signal compared to either a blank disc or a disc that had base oil (0 ppm standard) filtered through it. This demonstrates that organo-metallic compounds do not readily adhere to the carbon disc and that when certain elements are reported by the method they are presumably due to capture of particulates rather than due to absorption of chemical species by the disc.

Calibration: Spectrometer calibration is a challenge for the RFS method because multi-elemental particle standards do not exist. For the present, an absolute calibration is therefore impossible. However, if the RFS method is to be used as a trending technique, a calibration with verifiable standards will insure similar results from day to day and from instrument to instrument. Therefore, the normal RDE spectrometer calibration is used. This has the obvious advantage that an RDE spectrometer calibrated for routine measurements can also be used for RFS measurements. However, it should be understood that results for the RFS method do not report the actual parts per million value of the particles captured by the disk. The normal RDE calibration is done with organo-metallic standards. Therefore, for those impressed by semantics, RFS results can be thought of as "organo-metallic equivalent ppm".

Repeatability: Early RFS experiments displayed rather poor repeatability due primarily to poor sealing of the disc electrodes in the initial RFS filtration fixtures. These problems have been largely overcome. Repeatability, however, is not as good as achieved when analyzing organo-metallic standards. Repeatability of 10 to 20% relative standard deviation has been the experience to date, whereas repeatability of the normal RDE method is in the range of 3 to 5%.

Greater variation in results is expected when measuring particles dispersed in liquids rather than when measuring homogeneous solutions such as the oil standards. The sample must be shaken in a consistent fashion and must be withdrawn and delivered in a consistent manner. Anyone experienced with particle counters knows that attaining good repeatability requires some art.

Another factor that contributes to poor repeatability is the rather inconsistent porosity of the disc electrodes. Changes in porosity of the disc electrodes affects filtration efficiency. It has been found that if a data set of the same sample run repeatedly is normalized by the filtration time through the disc, the repeatability improves.

Nevertheless, the repeatability achieved by the RFS method is acceptable for predictive maintenance purposes. Identification of abnormal samples is straightforward since abnormal wear modes or contaminated oils usually have a great many more particles than samples from normally operating machines.

Application to a Predictive Maintenance Program: There are now at least three commercial laboratories and 15 or so industrial laboratories using the RFS method to overcome the large particle shortcomings of conventional spectroscopy. Two measurements are made for each oil sample. The first spectrometric measurement is done in a conventional manner giving the elemental concentration for dissolved material and very small particles. The second measurement is done by RFS which gives the elemental concentration for large particles. These two measurements, taken together, give an improved diagnostic capability compared to that previously available on a routine basis.

Applications example: A steam turbine oil sample gave either zero or one ppm readings for all elements measured by inductively coupled plasma (ICP) spectroscopy. [2] The RFS readings had a high silicon reading of 117 ppm., see Table II.

Particle count data was also elevated confirming the presence of large particles in the sample, see Table III.

Table II, Analysis of Steam Turbine Oil Sample with ICP and RFS Techniques

	Cr	Al	Ni	Ti	Ag	Cu	Sn	Mo	V	Pb	Mg	Fe	P	Si	Zn
ICP	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
RFS	1	1	1	0	0	3	3	0	0	1	2	19	7	117	0

Table III, Particle Count Data for Steam Turbine Oil Sample

	Size range	# of Particles
Particles per ml	5-10	2691
(μm range)	15-25	36
	25-50	28
	50-100	34
	>100	9
ISO >5 2798		*ISO >15 107*

Table IV, Ferrographic Results for Steam Turbine Oil Sample

Wear Particle Types	
Normal Rubbing	3
Severe Wear	2
Cutting Wear	0
Fatigue Particles	0
Laminar Particles	0
Spheres	0
Dark Metallo-oxide	3
Red Oxide	1
Corrosive	0
Nonferrous metal	0
Nonmetallic inorganic	6
Organic	0
Nonmetallic amorphous	0
Friction Polymer	0
Fibers	2

0 = None
2 = Trace
5 = Moderate
9 = Heavy

A ferrogram prepared from this sample showed a moderate amount of rather large non-metallic crystalline particles, see Table IV. These were in the 50 μm size range. A small amount of ferrous normal rubbing wear and severe wear particles were also present on the ferrogram. This example illustrates the weakness of conventional ICP spectroscopy to detect particulate contamination. Similar results can be expected when comparing the RFS data to analyses obtained with the RDE technique. It also demonstrates the usefulness of RFS to identify the composition of the particulate contaminants.

Discussion: For many years, the condition monitoring of critical machinery by oil analysis has often included investigation of whether large wear or contaminant particles were present in the sample. Some of the methods available were 1) preparation of a ferrogram, 2) particle counting, either automatic or filter patch testing, or 3) screening for large ferrous particles by direct reading (DR) ferrography or by other ferrous debris monitoring instruments. RFS is much faster, and therefore more economical to perform than preparation and examination of a ferrogram.

RFS is not a substitute for ferrography, but may be considered a screening test to determine when to prepare a ferrogram. RFS data complements ferrography in that particle composition is more adequately determined. Examination of the size, shape, surface characteristics and composition of the particles on a ferrogram allows identification of the wear mode from whence the sample came.

RFS can also be used as a screening test for when to perform particle counting. RFS complements particle counting by providing data on elemental composition. RFS may substitute for ferrous debris monitoring instruments in that RFS is sensitive not only to large ferrous particles but also to large particles of other composition.

Conclusion: The RFS method offers a new, sensitive screening test for large particles in used oil samples. It offers the important advantage of quantitative and qualitative determination of particulates in the oil sample. It has become an important additional test for oil analysis as it provides additional new information on the condition of a machine.

References

1) Anderson, D.P., U.S. Patent 5,408.306, 18 April 1995, "Rotating Disk Electrode Method and Apparatus for Multi-Elemental Determination of Particles in Used Oil"

2) Anderson, D.P., "Rotrode Filter Spectroscopy - A Method for the Multi-elemental Determination of the Concentration of Large Particles in Used Lubricating Oil", P/PM Technology, Sept./Oct. 1992, pp. 88-89.

OIL/WEAR PARTICLE ANALYSIS A PREDICTIVE MAINTENANCE TOOL

By Ray Dalley
PREDICTION
9555 Rockside Road
Cleveland, Ohio 44125

Abstract: Wear particle analysis and Ferrography in particular is an effective means to identify and respond to maintenance needs. The development of this technology includes image analysis, on-line sensors, Ferrogram Scanners, portable screening tools, automated oil analysis screening tools, electronic transfer of evaluation results, and artificial intelligence.

Wear is the inevitable consequence of surface contact between machine parts such as shafts, bearings, gears, and bushing...even in properly lubricated systems. Equipment life expectancies, safety factors, performance ratings and maintenance recommendations are predicated on normally occurring wear. However, such factors as design complexity, unit size, intricate assembly configurations, and variations in operating conditions and environments can make maintenance or repair needs (ordinary or emergency) difficult to evaluate or detect without taking equipment out of service.
Modern integrated and automated high speed machine systems make any interval of down time costly and non-productive. This is why non-interceptive diagnostic techniques such spectrometric oil analysis, vibration analysis, motor current analysis, and ferrography (wear particle analysis) are increasingly being applied in the power, process, semiconductor and manufacturing industries. Machine designers and builders are increasingly using wear analysis as a realistic criterion for improvements in products such as compressors, gears, bearings and turbine components. This paper will describe the wear particle analysis technology and it's effect working with other predictive maintenance tools within industry.

Key Words: Ferrography; predictive maintenance; Used Oil Analysis; Ferrogram Scanner; Passport System V; Wear Particle Types; Fourier Transform Infrared Spectrometer; Viscosity

Wear Particle Analysis/Ferrography

Ferrography is a technique that provides microscopic examination and analysis of wear particles separated from all type of fluids. Developed in the mid 1970's as a predictive maintenance technique, it was initially used to magnetically precipitate ferrous wear particles from lubricating oils.

- This technique was used successfully to monitor the condition of military aircraft engines, gear boxes, and transmissions. That success has prompted the development of other applications, including modification of the method to precipitate non-magnetic particles from lubricants, quantifying wear particles

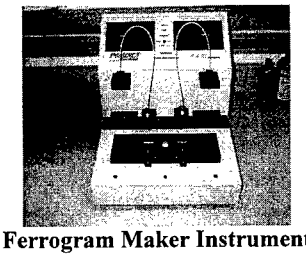
on a glass substrate (Ferrogram) and the refinement of our grease solvent utilized in heavy industry today.

- Three of the major types of equipment used in wear particle analysis are the Direct-Reading (DR) Ferrograph, the Analytical Ferrograph System and the Ferrogram Scanner.

Direct Reading (DR) Ferrograph: The DR Ferrograph Monitor is a trending tool that permits condition monitoring through examination of fluid samples on a scheduled, periodic basis. A compact, portable instrument that is easily operated even by a non-technical personnel, the DR Ferrograph quantitatively measures the concentration of ferrous wear particles in a lubricating or hydraulic oil. The DR Ferrograph provides for analysis of a fluid sample by precipitating particles onto the bottom of a glass tube that is subjected to a strong magnetic field. Fiber optic bundles direct light through the glass tube at two locations where large and small particles are deposited by the permanent magnet. At the onset of the test, before particles begin to precipitate, the instrument is automatically "zeroed" with a microprocessor chip as the light passes through the oil to adjust for its opacity. The light is reduced in relation to the number of particles deposited in the glass tube, and this reduction is monitored and displayed on a LCD panel. Two sets of readings are obtained: one for Direct Large >5 microns (DL) and one for Direct Small <5 microns (DS) particles. Wear Particle Concentration is derived by adding DL + DS divided by the volume of sample, establishing a machine wear trend baseline.

Machines starting service go through a wearing in process, during which the quantity of large particles quickly increases and then settles to an equilibrium concentration during normal running conditions. A key aspect of ferrography is that machines wearing abnormally will produce unusually large amounts of wear particles indicating excessive wear condition by the DR Ferrograph in WPC readings. If WPC readings are beyond the normal trend a Ferrogram sample slide is made with the fluid for examination by optical microscopy.

The Analytical Ferrograph: Additional information about a wear sample, can be obtained with the Analytical Ferrograph system, instruments that can provide a permanent record of the sample, as well as analytical information. The Analytical Ferrograph is used to prepare a Ferrogram -- a fixed slide of wear particles for microscopic examination and photographic documentation. The Ferrogram is an important predictive tool, since it provides an identification of the characteristic wear pattern of specific pieces of equipment. After the particles have deposited on the Ferrogram, a wash is used to flush away the oil or water-based lubricant. After the wash fluid evaporates, the wear particles remain permanently attached to the glass substrate and are ready for microscopic examination.



Ferrogram Maker Instrument

The Microscope: Ferrograms are typically examined under a microscope that combines the features of a biological and metallurgical microscope. Such equipment utilizes both reflected and transmitted light sources, which may be used simultaneously. Green, red, and polarized filters are also used to distinguish the size, composition, shape and texture of both metallic and non-metallic particles.

Ferrogram Scanner: The Ferrogram Scanner is a fairly a new patented instrument designed to scan ferrograms and quantify the analytical data; thus offering a new set of trending parameters comparing the results to a laboratory database of over 750,000 samples. This instrument uses a charged couple device (CCD) and the processing technology of a single board computer. By using a sensor array, the device scans the ferrogram made by the Ferrogram Maker in less than 20 seconds and generates standard output values that correspond to wear mechanism. The data output is manipulated with a software package controlled by a Windows based user interface. The user simply places the ferrogram on a slider and clicks on the computer interface. A motorized stage pulls the ferrogram into the instrument allowing the wear particle pattern to be scanned. The software then reports the wear levels and changes in the condition of the component. The program prompts the user through several graphing and condition analysis options, providing an overall recommendation.

Types of Wear Particles: There is six basics wear particle types generated through the wear process. These include ferrous and nonferrous particles a comprises:

1.Normal Rubbing Wear: Normal rubbing wear particles are generated as the result of normal sliding wear in a machine and result from exfoliation of parts of the shear mixed layer. Rubbing wear particles consist of flat platelets, generally 5 microns or smaller, although they may range up to 15 microns depending on equipment application. There should be little or no visible texturing of the surface and the thickness should be one micron or less.

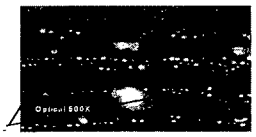
2.Cutting Wear Particles: Cutting wear particles are generated as a result of one surface penetrating another. There are two ways of generating this effect.

- A relatively hard component can become misaligned or fractured, resulting in hard sharp edge penetrating a softer surface. Particles generated this way is

generally coarse and large, averaging 2 to 5 microns wide and 25 microns to 100 microns long.

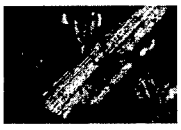
- Hard abrasive particles in the lubrication system, either as contaminants such as sand or wear debris from another part of the system, may become embedded in a soft wear surface (two body abrasion) such as a lead/tin alloy bearing. The abrasive particles protrude from the soft surface and penetrate the opposing wear surface. The maximum size of cutting wear particles generated in this way is proportional to the size of the abrasive particles in the lubricant. Very fine wire-like particles can be generated with thickness as low as .25 microns. Occasionally small particles, about 5 microns long by 25 microns thick, may be generated due to the presence of hard inclusions in one of the wearing surfaces.
- Cutting wear particles are abnormal. Their presence and quantity should be carefully monitored. If the majority of cutting wear particles in a system are around a few micrometers long and a fraction of a micrometer wide, the presence of particulate contaminants should be suspected. If a system shows increased quantities of large (50 micrometers long) cutting wear particles, a component failure is potentially imminent.

3.Spherical Particles: These particles are generated in the bearing cracks. If generated, their presence gives an improved warning of impending trouble as they are detectable before any actual spalling occurs. Rolling bearing fatigue is not the only source of spherical metallic particles. They are known to be generated by cavitation erosion and more importantly by welding or grinding processes. Spheres produced in fatigue cracks may be differentiated from those produced by other mechanisms through their size distribution. Rolling fatigue generates few spheres over 5 microns in diameter while the spheres generated by welding, grinding, and erosion are frequently over 10 microns in diameter.



Spheres : fatigue spalling eminent

4.Severe Sliding: Severe sliding wear particles are identified by parallel striations on their surfaces. They are generally larger than 15 microns, with the length-to-with thickness ratio falling between 5 and 30 microns. Severe sliding wear particles sometimes show evidence of temper colors, which may change the appearance of the particle after heat treatment.



Severe Sliding Wear

5.Bearing Wear Particle: These distinct particle types have been associated with rolling bearing fatigue:

- **Fatigue Spall Particles** constitute actual removal from the metal surface when a pit or a crack is propagated. These particles reach a maximum size of 100 microns during the microspalling process. Fatigue Spalls are generally are flat with a major dimensions-to-thickness ratio of 10 to 1. They have a smooth surface and a random, irregularly shape circumference.
- **Laminar Particles** are very thin free metal particles with frequent occurrence of holes. They range between 20 and 50 microns in major dimension with a thickness ratio of 30:1. These particles are formed by the passage of a wear particle through a rolling contact. Laminar particles may be generated throughout the life of a bearing, but at the onset of fatigue spalling, the quantity generated increases. An increasing quantity of laminar particles in addition to spherical wear is indicative of rolling-bearing fatigue microcracks.

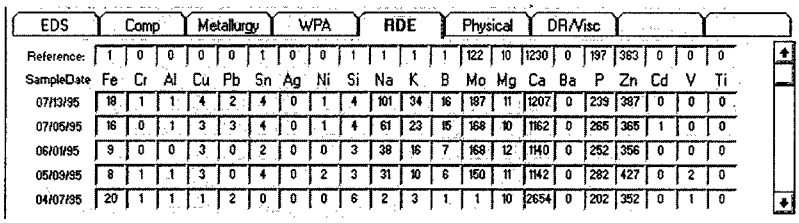
6.Gear Wear Two types of wear have been associated with gear wear:

- **Pitch Line Fatigue Particles** from a gear pitch line have much in common with a rolling-element bearing fatigue particles. They generally have a smooth surface and are frequently irregularly shaped. Depending on the gear design, the particles usually have a major dimension-to-thickness ratio between 4:1 and 10:1. The chunkier particles result from tensile stresses on the gear surface causing the fatigue cracks to propagate deeper into the gear tooth prior to spalling.
- **Scuffing or Scoring Particles** is caused by too high a load and/or speed. The particles tend to have a rough surface and jagged circumference. Even small particles may be discerned from rubbing wear by these characteristics. Some of the large particles have striations on their surface indicating a sliding contact. Because of the thermal nature of scuffing, quantities of oxide are usually present and some of the particles may show evidence of partial oxidation, that is, tan or blue temper colors.

Many other particle types are also present and generally describe particle morphology or origin such as chunk, black oxide, red oxide, corrosive, etc. In addition to ferrous and non-ferrous, contaminant particles can also be present and may include: Sand and Dirt, Fibers, Friction polymers, and Contaminant spheres.

Contaminant particles are generally considered the single most significant cause of abnormal component wear. The wear initiated by contaminants generally induces the formation of larger particles, with the formation rate being dependent on the filtration efficiency of the system. In fact, once a particle is generated and moves with the lubricant, it is technically a contaminant.

Used Oil Analysis: Used Oil Analysis (UOA) techniques include a light-emission spectrometry test for dissolved elements in the oil (particles generally no greater than 5-7 microns in size), Fourier transform infrared spectrometer analysis (FTIR) for physical and chemical changes in the oil, Kinematic Viscosity tests to determine oil viscosity degradation, various chemical titration's to determine depletion of alkaline reserve or build up of oil acidity (TAN & TBN), and chemical or distillation processes to determine water content.



	Fe	Cr	Al	Cu	Pb	Sn	Ag	Ni	Si	Na	K	B	Mo	Mg	Ca	Ba	P	Zn	Cd	V	Ti
Reference:	1	0	0	0	0	1	0	0	1	1	1	1	122	10	1230	0	197	363	0	0	0
SampleDate																					
07/13/95	18	1	1	4	2	4	0	1	4	101	34	16	187	11	1207	0	239	387	0	0	0
07/05/95	16	0	1	3	3	4	0	1	4	61	23	15	168	10	1162	0	265	365	1	0	0
06/01/95	9	0	0	3	0	2	0	0	3	38	16	7	169	12	1140	0	252	356	0	0	0
05/09/95	8	1	1	3	0	4	0	2	3	31	10	6	150	11	1142	0	282	427	0	2	0
04/07/95	20	1	1	1	2	0	0	0	6	2	3	1	1	10	2654	0	202	352	0	1	0

Used Oil Analysis Computer Window

Spectrometric Elemental Analysis: Spectrometric methods include atomic absorption (AA), atomic emission spectroscopy (AES), inductively coupled plasma emission spectrometer(ICP). Of these methods, atomic emission spectroscopy and induced couple plasma which rely on the detection of light emitted by the elements, are the most popular because of cost, speed, and other factors.

Fourier Transform Infrared Spectrometer: FTIR is widely used to determine water and coolant contamination of the lubricant, as well as to identify and monitor the depletion of additives and the buildup of oxidation products. A differential spectrum can be obtained by subtracting the spectrum of the new lubricant from that of the used lubricant to clearly reveal the area of change.

Viscosity: Probably the most important single property of a lubricating is its viscosity. It is a factor in the formation of lubricating films under both thick and thin film conditions; it affects heat generation in bearings, cylinders, and gears; it governs the sealing effect of the oil and the rate of consumption or loss; and it determines the ease with which machines may be started under cold conditions. For any piece of equipment, the first essential for satisfactory results is to use an oil of proper viscosity to meet the operating conditions. A decrease in viscosity may indicate contamination with a solvent or fuel or with a lower grade viscosity oil. An increase may indicate lube oxidation or contamination with a higher grade viscosity.

Total Acid Number: Acidity indicates the extent of oxidation of a lubricant and its ability to neutralize acids from exterior sources such as combustion gases. The acidity of lubricants is measured by the amount of potassium hydroxide required for neutralization (mg KOH/g) and the resultant number is called the TAN (total acid number). The additives in most new oils contribute a certain TAN or acidity, therefore, it is critical to

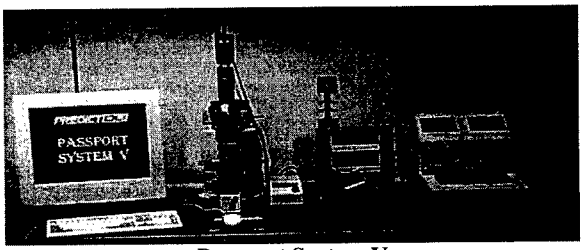
determine and monitor changes from new oil reference. An increase in TAN may indicate lube oxidation or contamination with an acidic product. A severely degraded lubricant indicated by a high TAN may be very corrosive.

Water Testing by Karl Fischer: This test produces iodine when electricity is conducted across a mesh screen. The electrical current needed to create iodine and remove existing water is measured and converted to parts per million (ppm). Quantification of water contamination - water in a lubricant not only promotes corrosion and oxidation, but also may form an emulsion having the appearance of a soft sludge.

Passport System V Software and Instruments

The combination and enhancement of WPA and UOA within the past few years have been oriented towards managing a predictive maintenance program efficiently with the advent of software and high tech tools. Of the recent development of our Passport System V software and instrumentation allows the user to incorporate all different types of predictive maintenance tools with a customized approach. The Passport System V is sophisticated, yet simple to use, state of the art data management and report writing tool, which provides users the most advanced capability available for computerized storage, comparison data, and evaluation of lube and wear data. The software design makes the creation of tables, charts, digitized pictures, drawings, and qualitative reports, previously produced manually, a faster and easier task, with more accurate results.

The Passport System V incorporates a video camera to capture and transmit the particle image magnified on the microscope to a personal computer. The data management features enable the technician to rapidly prepare a report and compare the current machine condition with the previous analysis history. The computer screen act as a regular display for report writing or reviewing information, while another part of the screen high resolution images are display from the microscope or from earlier reports, or pictures from the Wear Particle Atlas. With these combined features and having predictive maintenance information at your fingertips allows the technician to provide a comprehensive report with quality condition monitoring recommendations.



Passport System V

The **Passport System V** is a significant enhancement to predictive maintenance monitoring. The data management capability is obviously valuable to industry today. The technology to capture, digitize, transfer, and store images has progressed rapidly such that the quality is now equal or even better than still photography. The system enables the user to establish their own predictive maintenance reference guides for their machinery. The Modem/Internet e-mail system enables the technician to communicate with distant sister facilities' sites and with other technicians including our own experts. Perhaps most exciting of all is the prospect of developing an artificial intelligence base for diagnosis and decision making.

Diesel Applications

Wear particle analysis for diesel engines has been used in conjunction with other test methods, usually spectrometric analysis with superior results. As in other oil lubricated equipment, wear is an indicated by increasing amounts of particles and by changes in particle size distribution, composition, and morphology. The effects of engine operating conditions on the wear of cylinders' liners, piston rings, and crankshaft main bearings have been successfully observed by both wear particle and used oil analysis.
For diesel engines, heat treatment of the ferrograms distinguishes temper colors between low alloy (crankshafts) and cast iron (piston rings and cylinder liners), depending, of course, on the specific engine metallurgy. Although ferrous particles are primarily analyzed, other particles such as lead may be partially retained and has been used to follow main bearing wear. Normal Oil samples & ferrograms from diesel engines generally show only small rubbing wear particles and low iron levels on the spectrometric analysis test. A light deposit of corrosive wear debris at the ferrogram exit is typical. Diesels are exposed to acid conditions caused primarily by sulfur-containing fuels. In the United States this is becoming less of a problem, due to environmental regulations. In any case, the TAN testing process would confirm the acidity of the diesel lubricant.
Common wear problems in diesels are bore polishing, in which the cylinder wall is polished in spots to a mirror finish, and ring wear. Both of these problems are associated with piston deposits to some degree. This wear mechanism results in an increase in wear debris and that is detectable by both wear particle analysis and spectrometer.

Aircraft Gas Turbines

Aircraft and aircraft-derivative jet engines are subject to various failure mechanisms. Some of these failure modes proceed very rapidly, whereas others can be detected hundreds of operating hours before a shutdown condition is reached. Most failures of gas turbines occur in gas path. Gas-path failures frequently, but not always, cause an increase in wear particle size and concentration in the oil system, probably due to the transmittal of imbalance forces to turbine bearings and other oil-wetted parts. The resulting bearing or gear wear is then detected by both Used Oil Analysis and Wear Particle analysis.
Determining the exact source of wear problem can be difficult in a gas turbine because of complexity of the oil-wetted path. Typically several cavities, housing bearings, or gears will be force lubricated through individual return lines connected to a tank from which

the oil is pumped (at a high rate), then pass through a filter and heat exchanger, and the cycle repeated. Magnetic chip detectors or magnetic plugs are often installed in the return lines from various engine parts. These can help to pinpoint the source of generation in cases where particle metallurgy, as determined by heat-treating ferrograms, is similar for various engine parts. However, chip detectors will not give a warning until the wear situation is so severe that extremely large particles are being generated. By this time, the opportunity for predictive maintenance may be lost. Other analytical techniques, such as vibration analysis, may help to pinpoint the part in distress utilizing an expert system software that provides recommendations for action. In any case, predictive maintenance tools integrated together offer the maintenance engineer the best decision making tool.

Conclusion

The benefit of automation is in the use computer programs and emerging software technologies of artificial intelligence to assist in determining when to remove equipment from service for maintenance. For example, an advanced system which integrates emerging technologies in vibration, motor current analysis, thermography, ultrasonic, electronics, microprocessing, graphics, and data management could regularly sample a number of machines from a sampling device, compare the samples to previous samples for trend information (along with other Data parameters), make the decision to schedule the machine for maintenance, generate a work order for the maintenance team and send a purchase/work order to accounting for needed repair parts.

The maintenance manager/engineer could have almost instantaneous reports on the condition of each machine, along with a dollar figure indicating the optimal dates for shutdown and other maintenance requirements.

Technology advances oriented toward maintaining and incorporating all production data serve as an efficient assessment of manufacturing equipment. Companies as we know it today, can ill afford any shutdowns what so ever due to a tremendous amount of re-engineering or down-sizing occurring worldwide. Therefore, *predictive maintenance* tools working in conjunction with *production efficiency*, analyzed through a *cash flow model* are the decision making tools of today and tomorrow.

Acknowledgments

Wear particle analysis and Used Oil analysis information were extracted from the wear particle atlas and extensive experience of Predict/DLI employees. Other contributors to the preparation of this technical paper were Dr. Regan Silvestri, Robert Shinn, Vernon Westcott, and Bill Hoskins. A special thank you to my good friend Dan Anderson for his support and constant love for life.

Contact Raymond Dalley, Predict/DLI, 9555 Rockside Road #350, Cleveland, OH 44125; (216) 642-3223

MFPT

SIGNAL PROCESSING

CoChairs: **Kam Ng**
Office of Naval Research

Young S. Shin
Naval Postgraduate School

INTEGRATED LIFE MEASUREMENT AND DIAGNOSTICS FOR GAS TURBINE ENGINES

Michael J. Roemer
Stress Technology Incorporated
1800 Brighton-Henrietta Town Line Rd.
Rochester, New York 14623

Abstract: Accurate component life measurement and integrated diagnostics that reflect component design principles is critical to optimizing the structural integrity of gas turbine engine components once installed in an engine. State-of-the-art technologies such as "virtually sensing" currently unmeasured engine parameters such as turbine entry temperature (TET) are now being used in critical component lifing algorithms to enhance structural integrity assessments in a timely manner. In addition, complex finite-element and empirical models of structural and performance related engine areas can now be accessed in a real-time monitoring environment. Integration and implementation of these proven technologies presents a great opportunity to significantly enhance current engine health measurement capabilities and safely extend engine component life.

A strategy adopted by the USAF to develop a modular, comprehensive engine health monitoring system is presented. An R&D program whose ultimate aim is to develop and test an engine health monitor (EHM) capable of component life measurement, performance and mechanical diagnostics, sensor validation, trending, and anomaly detection is discussed. Engine data currently sensed and recorded for post flight processing will be analyzed in a continuous real-time mode. The measured data will be used as inputs to "virtual sensing" modules that will predict non-measured parameters needed in the life measurement module. Hence, life usage algorithms will determine critical component remaining life more accurately based on actual mission severity. For fault detection and accommodation, extensive knowledge of how a healthy engine operates under given conditions will be analyzed, and any deviation from this "normal" pattern of expected parameters will be detected and further analyzed. Faults resulting from sensor failure modes will be promptly isolated and more complex faults will be identified by pattern recognition schemes and fuzzy logic. The system under development is designed for the Rolls-Royce F405 engine (Adour) which is fitted to the Navy's T45 trainer, and a full-scale demonstration of the technology will ultimately be conducted on this engine.

Key Words: Gas Turbine Engine, Health Monitoring, Usage Monitoring, Diagnostics, Condition Monitoring, Artificial Intelligence, Life Measurement, Turbomachinery.

EHM System Architecture: The Engine Health Monitoring system under development consists of four modules that will perform sensor validation/recovery, trending, diagnostics, prognostics, and life usage functions. The four modules are named (1) data management, (2) health, (3) diagnostic and (4) engine life and operate simultaneously, transferring real-time, trended, diagnostic, and component life data between the modules as needed. The basic block diagram illustrating the functionality of each module is given below in Figure 1.

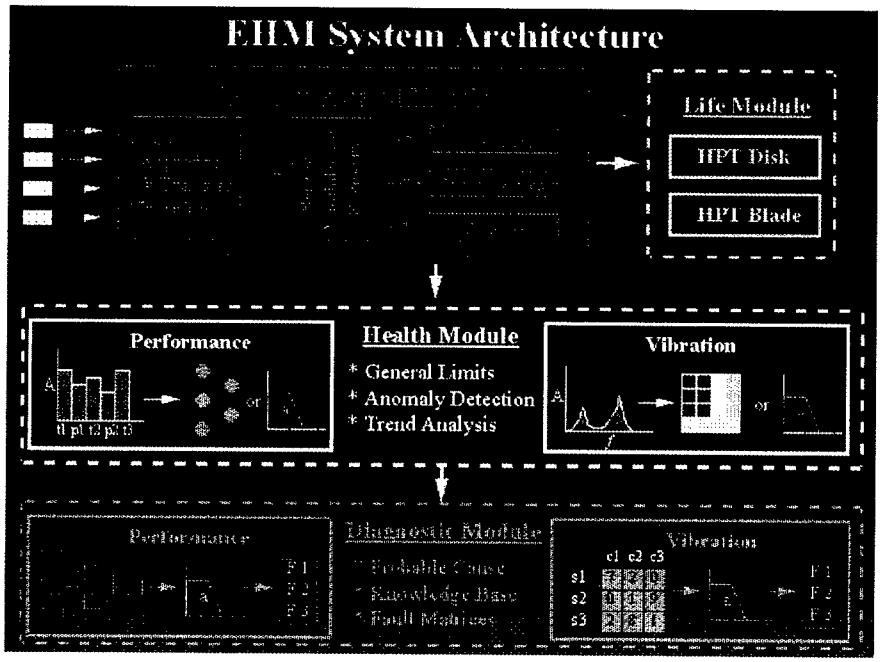


Figure 1 EHM System Architecture

Data Management Module: The first responsibility of the Data Management Module is to acquire the real-time, sensed engine data. Next, the module utilizes a neural network predictor architecture to perform several quality control checks to ensure the integrity of the sensed data. Details on the sensor validation scheme are given later in this paper. Once the data is thoroughly checked, additional engine parameters are predicted in real-time from the directly sensed engine parameters using a standard back propagation network architecture trained by "engine performance synthesis" models. Hence, currently unmeasured engine parameters such as turbine entry temperature (TET) can be predicted and utilized in real-time just as the directly sensed parameters. These "virtually sensed" engine parameters are then

used in the engine life module to obtain more accurate and realistic measures of hot component life parts. An engine duty monitor is also present in the data management module. Engine life is accumulated for both the HP turbine blades and disk based on LCF, creep, and thermal cycles and is presented as a fraction of total available life. The final function to be performed in this module is overall system data storage and retrieval. Raw sensed data, trended data, diagnosis results and engine life accumulation reports are all stored in this sub-module.

Sensor Validation and Recovery System

An advanced sensor validation scheme capable of detecting failed sensor hardware without sensor redundancy and during non-steady state monitoring conditions is a necessary "front end" to the EHM system. The developed approach utilizes neural networks and fuzzy logic to accomplish the desired goal. Neural networks are used to recognize the non-linear, inter-relationships between the different types of sensors used in a transient or steady-state measurement environment. Fuzzy logic is used to pre- and post-process the measurement data in order to determine general characteristics about the state of the process being monitored.

A block diagram of the sensor validation system architecture is given in Figure 2. The speed/power sensor data is first accepted by two parallel fuzzy logic modules. The first module determines the state of the speed/power condition (i.e increasing, decreasing, or steady-state) and the second verifies the validity of speed/power sensor itself. The output of the speed/power condition module triggers a particular neural network module that was specifically trained to know the sensor relationships for either increasing, decreasing or steady power output. Only one neural network module is triggered at a time, depending on the outcome of the prior fuzzy logic decisions. The sensor confidence values predicted by the neural networks are trended over time and passed through another fuzzy logic module to interpret the results. These extra steps are used to ensure that false alarms do not occur.

For the gas turbine engine application discussed in this paper, there are four primary performance related sensors (along with LP and HP rotor speeds) that are measured during turbine operation. These sensors include; fuel flow (Wf), HP compressor delivery pressure (P3), LP compressor delivery temperature (T2), and Jet Pipe temperature (T6). The outputs of the neural networks yield a confidence factor associated with the probability of a failed sensor. A confidence factor near one represents proper sensor operation, while a confidence factor near zero indicates a faulty sensor mode. A fuzzy logic module is used at the output of these neural networks to decide whether the sensor is good, bad, or somewhere in between. For instance, if a hard decision was utilized to alert the crew when a sensor confidence factor reached a level less than 0.80, false alarms would likely occur even though a sensor confidence factor of 0.78 might still indicate a properly working sensor.

SENSOR VALIDATION SYSTEM ARCHITECTURE
(NON-REDUNDANT; NON-STEADY STATE)

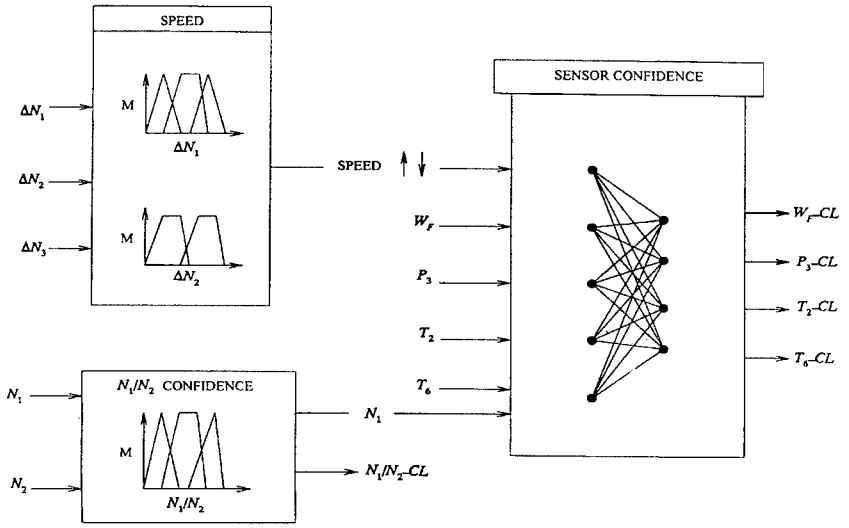


Figure 2 Sensor Validation Scheme

Health Module: The Health Module is dedicated to examining the real-time engine parameter data set and detecting any engine anomalies that exist with respect to the normal engine's operating signature. This function is performed for both the performance parameters and vibration data. First, the measured performance parameters are examined within pre-determined speed bands during the entire mission to ensure consistent and accurate performance patterns scans. When a set of acquired engine perfromance parameters trigger an anomaly being detected, the current real-time performance data is forwarded to the diagnostic module for detailed examination. Based on the trended data and fault pattern recognized, the severity and duration of the recognized fault can be determined.

Engine performance anomalies are detected by comparing normal "engine signatures" with a set of current measured engine data. Normal engine signatures are obtained for each engine during the standard production pass-off tests. Each measured parameter (T1, P1, T2, P3, T6, Wf, NL, NH) on the engine is plotted against the HP rotor speed to yield a "normal" range of operation for that specific engine. Any deviations from "fuzzy" bands encompassing these parameter plots will trigger the anomaly detection routines. The "fuzzy" bands are implemented with a dedicated fuzzy logic routine that contains membership functions that are specifically related to the particular engine being monitored.

In addition to the performance anomaly detection described above, vibration anomaly detection and diagnostics is also performed within the real-time monitoring environment. Vibration spectrums from the available engine installed accelerometers are gathered at particular engine operating speeds to form a "current" database of the engine's vibration characteristics. The speed ranges at which the spectrums are stored for the F405 engine are 80%, 85%, 90%, 95%, and 100% maximum NH rotor speed. Figure 3 is an illustration of the vibration anomaly detection and diagnostic scheme showing the spectrums gathered for the starboard tangential and port radial accelerometers.

Once the acceleration power spectral densities are organized into their respective bins, the peaks of the NL and NH rotor speeds are examined for overall amplitude and location (in terms of the % max NH speed) of the peaks. This information is used by the fuzzy logic diagnostic algorithms to determine if the vibration is coming from either the compressor or turbine and if it is associated with the HP or LP shaft. Once the basic location of the vibration source is determined, a more detailed diagnosis of the vibration problem is predicted. Currently, the system has been trained to recognize vibration fault patterns including; 1.) Unbalance/Misalignment, 2.) Rotor Rub and Mechanical Looseness, and 3.) Deteriorated Bearing Conditions.

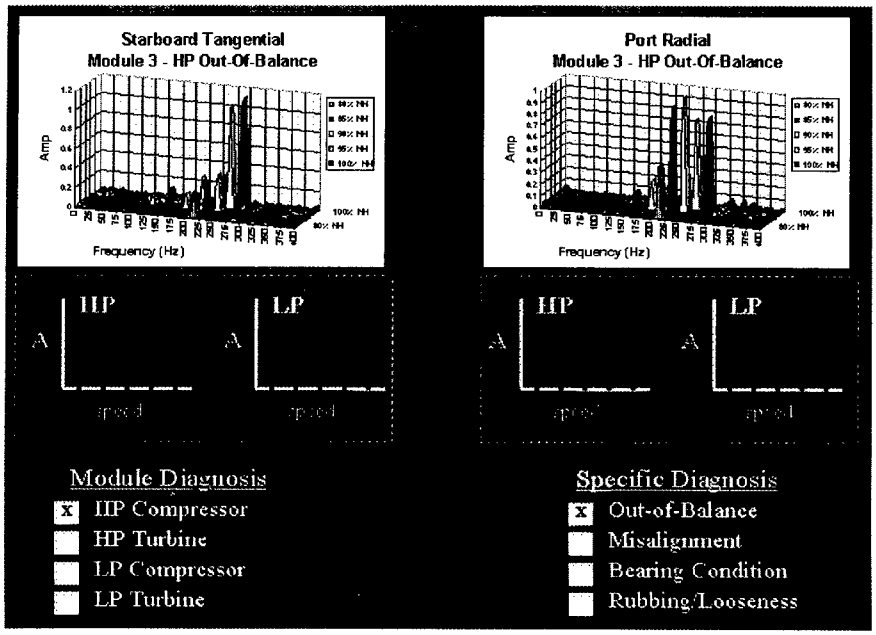


Figure 3 Vibration Anomaly Detection and Diagnostics Scheme

Diagnostic Module: Once an anomaly is detected and the current data is transferred to the Diagnostic Module, dedicated neural network fault classifiers determine the probable cause of the fault or degraded performance. A block diagram of the performance diagnostic procedure is given in Figure 4 shown below.

The underlying design of this model-based, fault diagnostic module lies in the determination of changing conditions appearing in engine sensory data due to the existence of particular faults. These observed changes are compared with the "normal" operating engine process to recognize error residuals. These residuals and associated patterns are then analyzed for failure detection and diagnosis by comparing them with known failure signatures associated with a failed component on the operating engine.

Engine fault signatures, which illustrate the effect of a failure on particular engine parameters, are generated from both historical engine data and engine synthesis models of the engine. In fact, a combination of these approaches is utilized in this EHM system. In the end, the failure diagnosis is accomplished with a neural network classifier to recognize the patterns of respective failure signatures. The performance diagnostic module in this program utilizes both self organizing neural network maps (to cluster and identify similar patterns) and trained network classifiers for specific problem diagnosis.

The real-time engine data for a particular speed-band is first compared with the corresponding "normal" operating patterns to determine measured parameter changes. These error patterns are then normalized with respect to the maximum measured parameter change and passed to a self organizing neural network map (Kohonen network) for initial pattern clustering. Figure 5 illustrates the results of Kohonen self-organizing map under noisy pattern conditions. Error pattern clustering was used as an initial diagnostic step because of its high robustness with respect to pattern noise. After the error pattern has been organized into a particular location on 10x10 interpretive Kohonen map, a trained back-propagation network classifies the coordinate location on the map into a specific diagnosis.

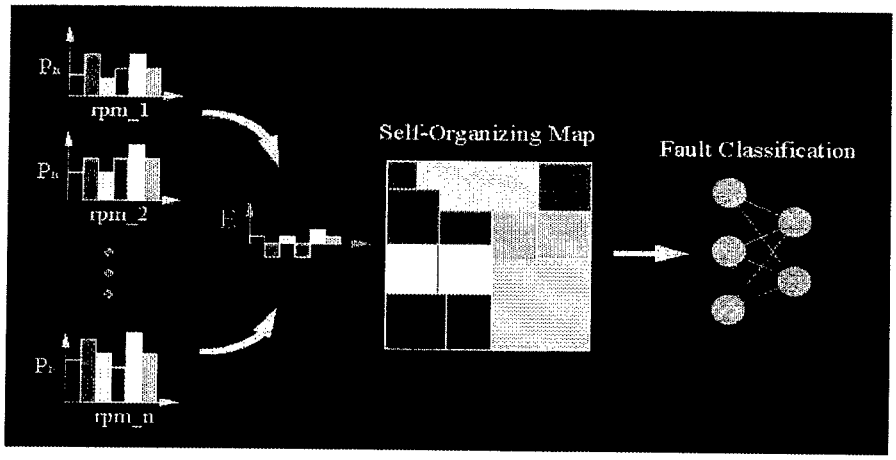


Figure 4 Performance Diagnostics Scheme

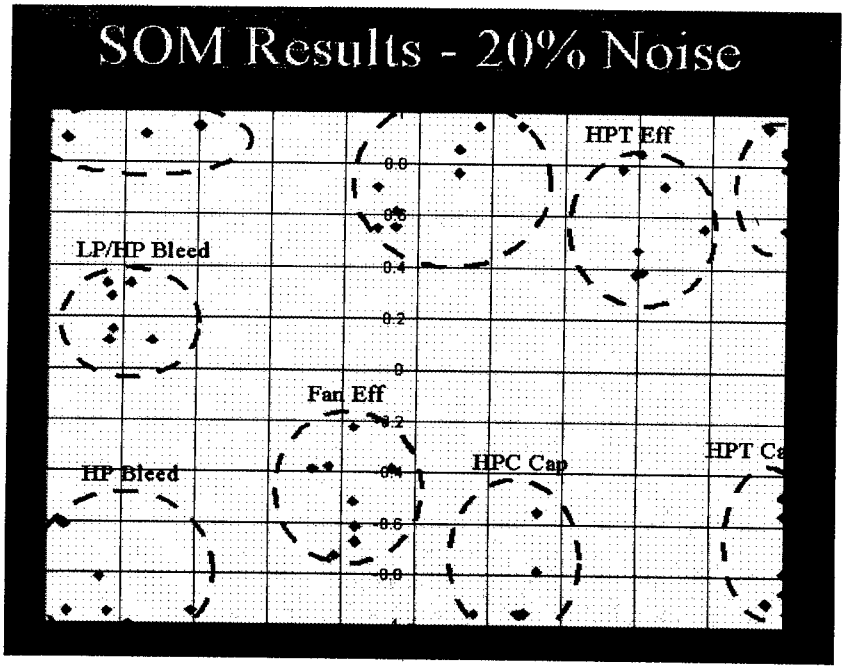


Figure 5 Kohonen Map Diagnostics

Engine Life Module: During the mission, data from the engine duty monitor within the Data Management Module is continuously being transferred to the Engine Life Module so that "mission representative" life may be accumulated. The life accumulation algorithm will include the engine specific parameters tracked throughout the entire mission. Hot section temperatures and stresses will be estimated in real-time based on a neural network architecture trained from engine synthesis models and BLADE-GT finite element models. Corresponding material creep and thermo-mechanical fatigue will therefore be more accurately determined based on actual mission severity. At the end of the mission, critical component life accumulated during that mission is reported and transferred back to the data management module for overall accumulation of each components life usage.

A significant advancement will be made in monitoring the life of HP turbine blades and disks by utilizing state-of-the-art BLADE-GT finite element models in real-time to predict the blade's critical stresses. The process for obtaining "virtual stress measurements" is accomplished by training neural networks with data that is generated by exercising the BLADE-GT model over a full range of engine operating conditions. This process has aleady been proven in several other applications with FE and engine synthesis models, with predictive virtual sensor results having an average error of less than 0.5%. Currently, the EHM system developed for the F405 engine on the Navy Trainer "virtually senses" SOT, T3, T4, and W1A.

Once the network is trained from the off-line model results, the network is ready to be accessed in real-time to produce the most accurate estimates of stresses and temperatures without actually having strain gauges or RTDs in place. This procedure has wide application in other areas of the engine where accurate models can be used to predict unmeasured parameters.

The discussed EHM system incorporates LCF, material temperature/creep, and thermo-mechanical fatigue aspects into a comprehensive damage accumulation algorithm. This represents a significant enhancement over the current life monitoring systems. Primarily, increased emphasis will be placed on "hot" components where these mechanisms play a large role in the failure process. The proposed approach of *individually tracking the total life of each critical component* will allow for a more representative and less conservative estimate of component life consumption.

The engine life module developed for the F405 engine concentrated on the HP turbine disk and blades Other components can easily be added to the system if the necessary life algorithms are available. Engine life usage algorithms for creep and low cycle fatigue of critical parts are already included in the Aircraft Data Recording system on the T45A Goshawk aircraft. The algorithms for the HP disk will be refined to use the more specific data available from the Data Management module to record accumulated life usage.

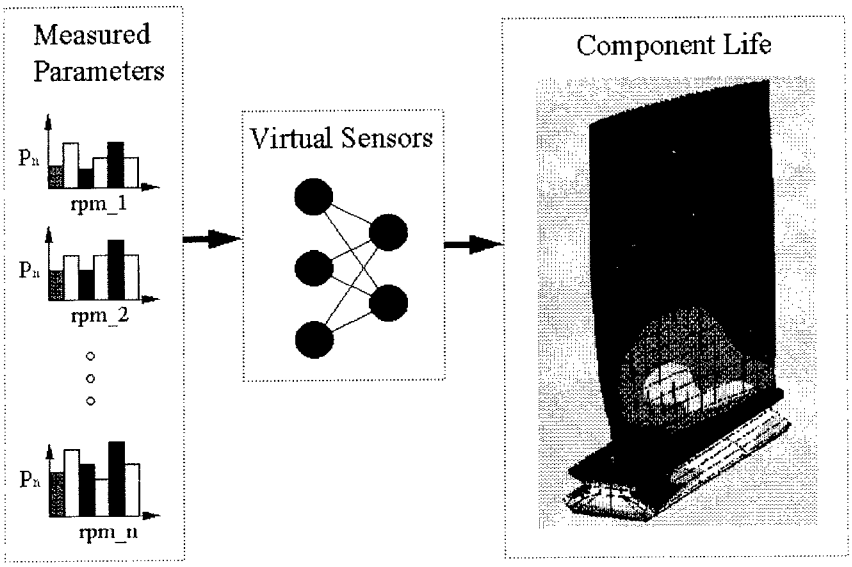


Figure 6 Engine Component Life Monitoring Scheme

EHM Ground Test System Software

A dedicated EHM software interface has been designed to aid in the development and testing of the real-time monitor. Specific user interface modules have been developed so that all monitoring aspects of the system can be examined during the testing stages of the program. Each of the engineering modules and corresponding software interfaces are discussed in this paper with examples of Adour engine data that have been processed by the EHM. The main "front end" EHM software screen is given in Figure 7.

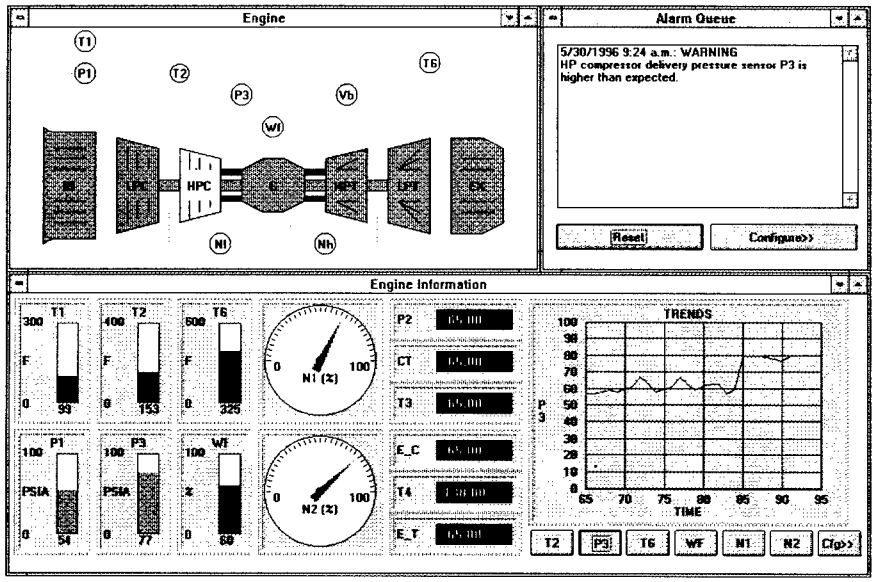


Figure 7 EHM Software "Main Screen"

Conclusions: With the application of artificial intelligence tc real-time engine diagnostics and more mission representative lifing algorithms, current engine health monitoring systems can be improved significantly. In particular, the incorporation of neural networks and fuzzy logic into the diagnostic process will yield great benefits in terms of processing speed, robustness, knowledge acquisition, and adaptability. A list of some important technical benefits of applying AI and advanced life prediction schemes to condition monitoring, diagnosis, and life usage are as follows:

1.) Enhanced ability to capture, organize, and utilize all relevant data, experience, and rules within a massively parallel, interconnected processing unit.

2.) Provides an automated procedure for incorporating data from several knowledge sources including; analytical FEM models, aerothermal performance models, empirical/trended test data, and heuristic (rules based) experience.

3.) Neural network system architectures are well suited for processing large quantities of non-linear, multidimensional, coupled parameters such as temperatures, pressures, etc. that occur within a gas turbine engine.

4.) Utilizing artificial intelligence results in a process that is less dependent on human experts and more automated to facilitate quick decision making.

5.) Incorporating more accurate information into the proposed life accumulation algorithms will provide a more accurate outlook on the remaining life of critical engine components.

References

1. Walker, N. D. and Wyatt-Mair, G. F., "*Sensor Signal Validation Using Analytical Redundancy for an Aluminum Cold Rolling Mill*", **Control Engineering Practice**, Vol. 3, No. 6, June 1995, pp.753-760.

2. McAvoy, T. J., "*Sensor Data Analysis Using Autoassociative Neural Nets*", **World Congress on Neural Networks**, Vol. 1, 1994, pp.161-166.

3. Holbert, K. E., Heger, A. S., and Alang-Rashid, N. K., "*Redundant Sensor Validation by using Fuzzy Logic*", **Nuclear Science and Engineering**, Vol. 118, No. 1, Sept. 1994, pp.54-64.

4. Harrison, P. R., and Harrison, P. A., "*Validating an Embedded Intelligent Sensor Control System*", **IEEE Expert**, Vol. 9, No. 3, June 1994, pp. 49-53.

5. Ahmadi-Echendu, J. E., and Hengjun, Zhu, "*Detecting Changes in the Condition of Process Instruments*", **IEEE Transactions on Instrumentation and Measurement**, Vol. 43, No. 2, April 1994, pp.355-358.

6. Lee, S. C., "*Sensor Value Validation Based on Systematic Exploration of the Sensor Redundancy for Fault Diagnosis*", **IEEE Transactions on Systems, Man and Cybernetics**, Vol. 24, No. 4, April 1994, pp. 594-605.

HYBRID REASONING TECHNIQUES FOR AUTOMATED FAULT CLASSIFICATION

Amulya K. Garga, Ph.D.
David L. Hall, Ph.D.

Applied Research Laboratory
The Pennsylvania State University
P.O. Box 30
State College, PA 16804-0030
Phone: +1 814-863-5841
Fax: +1 814-863-0673
Email: garga@psu.edu

Abstract: The detection, characterization, and classification of fault (or impending fault) conditions in complex machinery usually involves the use of either implicit reasoning techniques or explicit reasoning techniques. Implicit reasoning techniques, such as artificial neural networks (ANN) or pattern recognition techniques, transform observed data into a declaration of the status or health of a monitored system. Neural nets, for example, which must be trained with a priori data, encode a (generally) non-linear transformation via a set of weights associated with the network architecture. By contrast, explicit automated reasoning techniques, such as rule-based expert systems or logical template systems, encode explicit knowledge gained from maintenance personnel or system designers to allow interpretation of fault conditions based on the values of observed data. An automated reasoning technique that has been found useful is a hybrid approach which combines implicit and explicit automated reasoning techniques. In particular, a set of fuzzy logic rules are first developed to reason about fault conditions based on information obtained from knowledgeable maintenance personnel. Next, a neural network is trained to represent the rules. Finally, the neural net is supplied with examples of real data for fault and non-fault conditions and allowed to evolve or re-learn to adapt to the supplied data. The resulting hybrid neural network is more robust than could be obtained by simply training the neural network from sample data alone. Moreover, it is possible (by inspection) to derive rules from the neural network which are usually more concise and robust than the original rules. These features facilitate accurate assessment of fault conditions in a machine, which is essential for predicting the remaining useful life of the machine.

Key Words: Automated reasoning; condition-based maintenance; explicit reasoning; fault classification; fuzzy logic; hybrid reasoning techniques; implicit reasoning; neural networks

Introduction:

Condition-based maintenance (CBM) of mechanical systems has received increased attention recently. The goal of a CBM system is to *monitor* the operation of a complex mechanical equipment and provide the operator with an *accurate assessment* of the system's current health (state), and, if possible a *prediction* of the remaining useful life of the equipment [3, 6]. The success of such an effort depends on the development of improved robust sensors, real-time techniques for processing sensor data, fusion techniques for combining resulting refined information and micro and macro level models, and automated reasoning techniques to interpret the results of the previous analysis in the context of the operational environment and the task goals. CBM represents a migration from the traditional maintenance approaches where maintenance is either *restorative*, i.e. performed after a failure occurs, or *schedule-based*, i.e. performed after pre-determined intervals, which are typically based on worst-case mortality statistics [3]. Though the traditional approaches usually avoid catastrophic failure, they lead to unnecessary downtime, and either unacceptable risk (with restorative maintenance) or inefficient utilization of resources (with schedule-based maintenance). Crucial to CBM is the ability to monitor the machine's condition and accurately assess fault conditions. This requires monitoring of the machine's vibration, acoustic emission, temperature, and pressure, etc. However, to prevent information overload it is important that the sensor data be presented in a more refined and concise manner. This can be accomplished by *sensor data fusion* to remove redundant information and to combine imprecise information, in order to obtain distilled and accurate information. This is possible because different sensors provide information about different aspects of a complex system and data fusion utilizes the most reliable information from each sensor to provide more accurate information about the complex system than is possible to obtain with any one sensor [4]. But the value of the refined information is only realized when it is interpreted with respect to the machine's task or the mission scenario. This is one of the main goals of automated reasoning for CBM.

While automated reasoning research has been ongoing for several decades it is only recently that due attention is being focused on its application to large complex systems. As this happens it is becoming increasing clear that there is no single approach that can solve all the problems in creating an automated reasoning system for complex systems. Researchers have developed symbolic, connectionist, and statistical approaches for automated reasoning. There remain many challenges in this area and some of the challenges especially in the context of CBM are listed in the next section.

At the Pennsylvania State University a number of experiments are being conducted to develop techniques to diagnose and predict fault conditions for mechanical systems. In this paper we show how the combination of fuzzy rule-based and neural network approaches can lead to better performance for monitoring and fault classification. In the next section the challenges for an automated reasoning system are briefly described and some techniques that address those challenges are identified. Then, the hybrid approach for automated reasoning is presented and illustrated via an example application of an intelligent oil lubrication monitoring system. Finally, a summary of the new approach and outline future research and development areas are presented.

Challenges and Techniques for AR for CBM:

An intelligent system must monitor its components and systems, assess their health, and identify the implications of that assessment to the health of the system and to its mission. These requirements are very similar to those for the development of a CBM system [3]. Indeed, situation monitoring, its assessment, and predicting its implications, or **MAP**, would constitute intelligent behavior for a human, for an animal, and, for a machine.

Only recently has there been effort to develop automated reasoning systems for practical complex systems such as large machinery. Contemporary automated reasoning approaches grew out of the efforts of Hebb, McCulloch, and Pitts, on modeling the behavior of the brain in the 1940s, and the efforts of Rosenbaltt, Wiener, von Neumann, Minsky, Widrow and others, on making intelligent machines in the 1950s. But these approach did not succeed, largely due to the inadequacies of the technology of that time [2]. Over the next two decades, there grew are two main branches of the field that Minsky called Artificial Intelligence (AI), *symbolic* and *connectionist*. At the same time, a small but hardy group of researchers set out to capture two essential aspects of intelligence, *linguistic reasoning*, with Fuzzy Logic theory proposed by Zadeh, and *imprecise reasoning*, with the Demspter-Shafer Belief Theory [8]. Interestingly, symbolic AI camp primarily utilized probability theory and belief theory while connectionist AI camp utilized fuzzy logic theory and approximation theory.

In Table 1 above, a summary of some challenges faced in developing an automated reasoning system is given [5]. The approach described here suitably addresses the first five challenges summarized in Table 1, and is readily applicable to all but the last challenge. Its application to the last challenge would require some modification and a suitable control structure.

Table 1. Summary of Challenges to AR for CBM

Challenge to AR	*Summary*
Default reasoning	the ability to achieve a conclusion with "reasonable" assumptions and the absence of information
Explicit and Implicit Knowledge Representation	the ability to represent explicit knowledge (e.g. rules, templates, frames, etc.) and implicit knowledge (e.g., stored in neural nets)
Negative Reasoning	the ability to explicitly represent and reason with negative information, the implications of the absence of information
Truth Maintenance	a method to maintain uncertainty structure associated with the evolving conclusion
Practicality	an approach which leads to a finite, and potentially a real-time, implementation

Table 1. (cont'd) Summary of Challenges to AR for CBM

Hierarchical Knowledge Representation and Reasoning	the capability to reason about hierarchically organized systems, to exploit the natural partitioning of the problem
Multiple Time-Scale Causal Reasoning	the ability to reason with asynchronous information with very dissimilar rates of variation (time-scales)
Ordering of Extensions	a technique to deal with multiple, possibly conflicting, assumptions
Non-Monotonic Reasoning	the capability to retract or reverse a conclusion based on further information

A Hybrid Approach for AR for Fault Classification in CBM Systems:

A hybrid approach *combining fuzzy logic and rule-based systems*, to capture the expert's explicit knowledge, and *neural networks*, to provide a parsimonious representation of the knowledge base and for their adaptability is described here. The approach was developed for automating intelligent condition monitoring applications and provides an efficient implementation mechanism for distributed diagnostics.

Typical AR systems either utilize implicit knowledge or explicit knowledge about the complex system to be monitored. Usually, expert systems or rule-based systems employ explicit information acquired from experts while in neural network implicit information is stored in their interconnections of the neurons which are trained for with raw sensor data. The main limitations of rule-based systems are (a) the combinatorial explosion that results from the need to account for as many combinations of the input variables as possible and (b) the problem of inconsistent or conflicting rules which are hard to detect as the rule base becomes large. The first limitation is partly obviated with the use of fuzzy logic instead of symbolic logic and this has the pleasant side effect of reducing the severity of the second problem as well. Nevertheless, even fuzzy rule-based systems are not designed to handle both explicit and implicit knowledge simultaneously. While neural networks are able to readily capture the implicit knowledge hidden in the variation of sensor data, their main limitations are (a) it is usually very hard to explain the reason that the network made specific decisions and (b) training techniques usually take very long to converge and require trial and error and educated guesses to decide the network size, topology, and initialization. An advantage of explicit systems is that they can be developed with minimum prior test data. A disadvantage of such systems, however, is that they may not reflect the realities of the actual systems operating in realistic (as opposed to laboratory based) environment, i.e., the explicit system may be based on machinery specifications rather than on the actual conditions experienced during operation. It is clear that the two methods compliment each others advantages and limitations. It is natural, then, to combine the two approaches to exploit their advantages and to eliminate their disadvantages. This is precisely the basis for our approach.

The Hybrid Framework

We now provide an overview of the approach for automated fault classification in complex systems. First, a knowledge engineer elicits explicit rules from a domain expert about the system's operation. These are used to train a neural network to learn the rule set. The antecedents in the rules are fuzzy variables and, therefore, they take values from a finite discrete set, which can be mapped to any set that is isomorphic to a finite subset of the set of natural numbers. This important observation results in two consequences: the rule-base is finite, though its size is an exponential function of the number of unique variables and it allows the interpretation of the neural network, trained with such rule-set.

The acquired rules are used to train a layered feedforward neural network to capture the explicit knowledge. Finally, a much smaller rule set is derived from the neural network which represents the same explicit knowledge that was obtained from a domain expert. The neural network can be retrained on actual sensor data to refine the knowledge base and to incorporate implicit system knowledge. Rules can also be extracted from the resulting network. The notable features of the novel approach are its ability to *encapsulate explicit and implicit knowledge* and its *flexibility in allowing updates* of the knowledge base and *extraction of a parsimonious and consistent set of rules* from the network. The new hybrid approach uses a combination of fuzzy rule-base and a neural network knowledge representation methods. shown pictorially in Figure 1. In describing the approach the focus is on the CBM application of the inferring state of an internal (critical) part of a machine by monitoring the outputs of sensors in the machine. It is rare to be able to locate the sensor close to the hidden part and several sensors are utilized to get a reasonable assessment of the critical component's state. In the future, this may be less of a restriction when new nanosensors become available, which could be placed very close to, if not on, the critical parts [3, 6].

When using neural networks the key issues to address are:
(a) the number of inputs and outputs?
(b) the number of hidden neurons and their activation functions?
(c) the type of training/learning?
(d) the amount of training data required for adequately learning the knowledge base?

The main advantages of the new approach are:

- captures explicit knowledge via fuzzy logic rules,
- parsimonious representation of implicit knowledge,
- captures implicit knowledge from sensor inputs during operation,
- fast evaluation for real-time operation,
- adapts to particular machine or system and to operational conditions,
- extract rules from neural network to explain operation and refine knowledge,
- improve understanding of system by identifying key components affecting system operation, and
- the network permits the extraction of both positive and negative reasoning type rules - what is wrong vs. what isn't wrong [5], greatly helps in reducing the size of rule sets and search space.

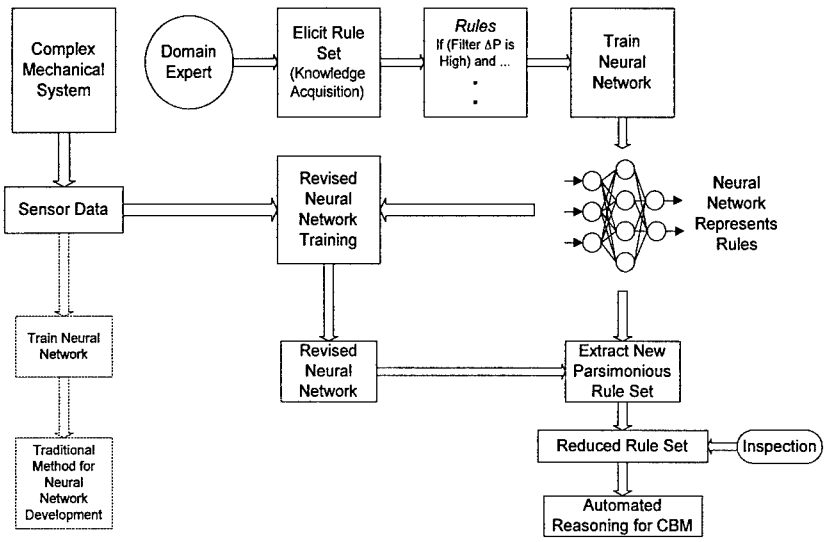


Figure 1: The hybrid implicit/explicit automated reasoning approach.

An Example Application - Oil Lubrication System Monitoring

Now an application is presented to demonstrate the new hybrid approach for automated reasoning. In an oil lubrication system it is often of considerable importance to assess if the oil filter is clogged or not. Typically this assessment is made by removing and visually inspecting the filter. This is not feasible when the machine is in operation and can cause significant down-time if done frequently. An alternative is proposed that can perform the assessment of the oil filter's condition on-line. An oil lubrication monitoring system would monitor observables from various sensors and use a knowledge-base, established with the help of an oil lubrication expert, to assess the filter's condition continuously. Such a system can be a valuable asset since it would provide timely assessment of the filter's condition and permit corrective or even preventive actions to be taken.

Sample Rules Acquired from Expert

In this example of an intelligent oil lubrication monitoring system, a set of explicit rules is built, for ascertaining whether the oil filter is clogged based on pressure, temperature, and flow measurements. The first step in the implementation of an expert system or, in general, a reasoning system is the establishment of a knowledge base. This is typically, done by querying a domain expert and eliciting rules that could be used in arriving at conclusions about the machine or complex system using facts about its present and past condition. The rules are usually cast as Horn-clauses, i.e., **If** $(x_1\ C_1\ x_2\ C_2\ \ldots\ x_n)$ **then** y,

where x_i are variables or facts and C_i are AND/OR connectives (i = 1, 2, ..., n). The expression in the parentheses after the If is called the premise or the antecedent and y constitutes the conclusion or the consequent. Note that the variables (facts) x_i can be positive or negative, i.e., they could signify the presence or absence of a condition or event. A more general representation would allow for multiple consequents, and could result in reduced number of rules by combining those rules that have identical antecedent.

For the purpose of illustrating the new approach an oil lubrication system was selected, whose health is important for most mechanical equipment. Three sensors were selected to measure differential pressure drop though the filter, *Filter* ΔP, *Temperature* close to the filter and can give an indication of viscosity change, and the oil flow which varies with engine speed, *Volume Flow*, and these variables are designated in the neural network by ΔP, T, and F, respectively. The inferred variables are *Filter Clog*, a boolean variable indicating, and *CF*, a confidence factor indicating confidence in the value of *Filter Clog*, which will be referred to concisely as C. Each sensor output was to be read as either Low, Normal, or High. An example rule from a rule set acquired from an domain expert is

> If (***Filter*** ΔP is **Low**) and (***Temperature*** is **Normal**) and (***Volume Flow*** is **High**) then (***Filter Clog*** is **No**) ***CF*** **0.9.**

A one-hidden layer feedforward neural network was selected and trained to learn the initial rule set, which consists of three layers neurons as shown in Figure 2. Note that there are no intra-layer connections, which is the defining characteristic of strictly feedforward networks. Each neuron (node) simply linearly combines all its inputs with their respective connection weights and evaluates the net result with its activation function. The activation function used was the sigmoid, $f(x)= 1/\left((1+\exp(-\alpha x)\right)$, where α controls the steepness of the function in its linear region. The network has three inputs (ΔP, T, F) and one output as necessitated by the selection of sensors and the desired variable to be inferred (C). Since the input layer of neurons just provides the inputs to the hidden layer it is often not counted when describing the network and the network with one hidden layer is also called a two-layer network.

The fact that the inputs were fuzzy variables which took values in {Low, Normal, High}, which we represented as {0, 0.5, 1}, permitted the use of standard neural network training. Specifically, this selection results in reasonable numerical results with the neural network training algorithms in neural networks toolbox of MATLAB, using the adaptive backpropagation of errors with momentum technique. The output C is boolean and it was represented as a continuous variable to allow faster training. Later it was thresholded to obtain a boolean output, with a value greater than 0.6 signifying clogged filter. The confidence factor CF was derived separately, as described below. The number of hidden units (7) was arrived upon after trying many different configurations. It is significant to note that only a part of the original rule-set was used to train the neural network and it responded correctly for all the rules. Such generalization capability is one of the attractive features of neural networks.

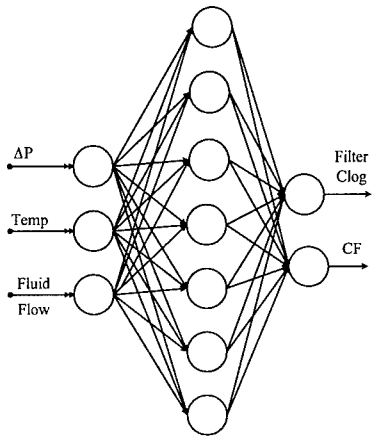


Figure 2: A neural network for the oil lubrication monitoring system.

Reduced Rule Set

A **reduced set of 4 fuzzy rules** was derived, which depicts a mere fraction of the size of the original rule-set. However, the reduced rule-set is *not a subset of the original set*. The rules in the reduced set are more complex, in the sense that the antecedent for a *typical new rule accounts for the antecedents of several of the old rules*. Furthermore, the new rules *utilize negative information to reduce the complexity* of the rule-set. It may be possible to produce even smaller rule-sets by using other rule extraction methods. An example of a rule resulting from the trained neural network is

> If (***Filter*** ΔP is **High**) and (***Temperature*** is **Not Low**) and (***Volume Flow*** is **Not High**) then (***Filter Clog*** is **Yes**) ***CF*** **0.81**

This rule accounts for and replaces four of the original rules. The confidence factor 0.81 was obtained by multiplying the confidence factors, for those original rules and is lower than for any of the four original rules. This is due to the specific expression (multiplication) used for obtaining the combined confidence factor. The issue of aggregating and propagating fuzzy information is a very active area of research and a neural approach is addressed in [9].

Conclusions and Future Work:

An reasoning approach for automated fault classification for condition-based maintenance systems was described. Logical rules based on the structure of the mechanical system and on engineering guidelines can help in assessing whether indications coming from multiple sensors are indicative of fault or not. Such a knowledge base can be refined and adapted with specific machinery data to reflect its operating behavior. The approach described is a hybrid approach both in the way it represents the it combines a rule-based knowledge representation and a neural network implementation. The rule-based approach provides for easy understanding of the system's operation and the neural network

implementation provides for an efficient and on-line operation. Furthermore, the neural network approach is amenable to update based on the operating condition of each machine. Interpretation of the network's knowledge after extracting rules from it should help identify which components or subsystems are more critical than others and help in the overall understanding and modeling of the machine's behavior. The neural network permits the extraction of both positive and negative reasoning type rules. It is possible to derive rules that utilize the information about what's wrong with the system and other rules that utilize the information about what's not wrong with the system. Moreover, this flexibility results in reducing the size of the resulting rule sets. This has favorable impact on the speed of operation of the reasoning system making it a viable candidate for real-time operation. Furthermore, the small size of the rule-set should permit its incorporation into small intelligent modules that are located at the machine level and this leads to local and distributed reasoning capability that could be crucial for a large and complex system such as a ship [1] or a helicopter.

The availability of good training data has been assumed in this paper. Good training data, especially transitional data which is crucial for developing machinery prognostic models, is difficult to come by. A mechanical diagnostics test bed was implemented under the MURI/IPD program to collect transitional data for gearbox faults. The test bed is described in [7], which also contains description of its design, the test-plan design, and safety and control mechanisms.

In closing, several areas of future research are identified that are relevant to the areas of CBM and automated reasoning.

1. Rule extraction from neural networks has been addressed only very recently. Adapting logic minimization techniques, function/mapping learning [9], geometric approaches [2],and region-growing approaches, are just some of the interesting avenues that can be pursued.
2. Other types of networks can be employed instead of strictly feedforward topology, such as recurrent networks. Other types of activation functions and other types of learning scheme can be used, e.g., radial-basis networks.
3. The fuzzy logic approach deserves further exploration to perform as much simplification of the original rule set as possible.
4. Fuzzy function learning can replace the neural network, for implicit reasoning, in the above approach and should be amenable to rule extraction. The area of fuzzy reasoning and the learning of fuzzy is still largely unexplored and will have an enormous impact on other areas, such as neural networks, automated modeling of nonlinear dynamical systems, and in the CBM automated reasoning systems.
5. Under the MURI/IPD program we are establishing data acquisition and archival standards so that researchers are able to optimally utilize the data that are collected. Properly collected and documented data lead to high quality research, and have largely a positive impact on the area of automated reasoning as well. Improved electronic documentation efforts are an important aspect of this approach and should benefit CBM programs.

References:

[1] Garga, A. K., "A Hybrid Implicit/Explicit Automated Reasoning Approach for Condition-Based Maintenance," *ASNE Intelligent Ships Symposium II*, Philadelphia, PA, November 25-26, 1996.

[2] Garga, A. K., *Neural Network Design and Training using Computational Geometry*, Ph.D. Thesis, Department of Electrical Engineering, The Pennsylvania State University, 1994.

[3] Hall D. L., "Concepts of condition-based maintenance for intelligent ships," *ASNE Intelligent Ships Symposium II*, Philadelphia, PA, November 25-26, 1996.

[4] Hall, D. L., *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Norwood, MA, 1992.

[5] Hall, D. L., R. J. Hansen, and D. C. Lang, "The negative information problem in mechanical diagnostics," *Proceedings of the International Gas Turbine Conference*, London, June 1996.

[6] Hansen, R. J., D. L. Hall, and S. K. Kurtz, "A new approach to the challenge of machinery prognostics," *International Gas Turbine and Aeroengine Congress and Exposition*, Netherlands, June 13-16, 1994. (Also in the *Transactions of the ASME*, **117**:320-325, April 1995.)

[7] Kozlowski, J. D. and C. S. Byington, "Mechanical diagnostics test bed for condition-based maintenance," *ASNE Intelligent Ships Symposium II*, Philadelphia, PA, November 25-26, 1996.

[8] Post, S. and A. P. Sage, "An overview of automated reasoning," *IEEE Transactions on Systems, Man, and Cybernetics*, **20**(1):202-224, January 1990.

[9] Stadter, P. A. and A. K. Garga, "A Neural Architecture for Fuzzy Classification with Application to Complex System Tracking," in review for the *International Conference on Neural Networks*, 1997.

Acknowledgments:

This work was supported by a grant from the Office of Naval Technology (the Multi-Disciplinary University Research Initiative for Integrated Predictive Diagnostics Grant no. N00014-95-1-10461). We would like to thank Derek Lang for several interesting discussions. We would also like to thank Carl Byington for providing the original rule-set for the oil lube system.

GEAR AND BEARING DIAGNOSTICS USING NEURAL NETWORK-BASED AMPLITUDE AND PHASE DEMODULATION

E.C. Larson, D.P. Wipf, and B.E. Parker, Jr.
BARRON ASSOCIATES, INC.
3046A Berkmar Drive
Charlottesville, Virginia 22901-1444

Abstract: This paper is concerned with diagnostic methods for gears and bearings, the goals of which are to detect incipient structural and metallurgical faults, characterize their nature and severity, and isolate them to particular components. Such information is invaluable for prognostic purposes. The methodology presented herein focuses on the analysis of vibration signals induced by gears and bearings and, specifically, on a means of extracting from such vibration patterns amplitude and phase modulation signals. Such modulation terms, we conjecture, are simple indicators of the severity of a particular type of localized component defect. We show that such amplitude and phase modulation information can be extracted using a neural network computational methodology that relies on nothing more than knowledge of the bearing geometry and the frequency tones at which a given type of defect will manifest itself. We present numerical simulation examples and show that the technique requires extremely few *a priori* assumptions.

Key Words: Condition-based maintenance, Machinery diagnostics, Health monitoring, Incipient faults, Gears, Bearings, Modulation, Polynomial neural networks

Introduction: The topics addressed herein pertain to the general problem of machinery *prognostics*, i.e., estimating the remaining useful life of machinery components based on knowledge of their present condition and assumptions about future usage. The ability to diagnose particular machine components and understand the implications of diagnostic information vis-à-vis future usage of the machine is of immense practical importance in industrial, commercial, and military applications. Component failures contribute to the majority of machine operation safety incidents and are responsible for a substantial fraction of the on-going maintenance costs accruing to machinery usage. This underscores the need for *condition-based maintenance* (CBM), wherein knowledge regarding the *condition* of machinery, rather than fixed time intervals, is used to schedule maintenance, thus averting unnecessary inspections.

The quality of the diagnostic information acquired (i.e., fault detection, type and severity characterization, and component isolation) plays a paramount role in determining the efficacy of prognostication. Use of inaccurate diagnostic information in, say, a prognostic decisional algorithm driving a CBM policy for a fleet of helicopters may result in ill-fated missions. Prompt and accurate detection of structural and metallurgical faults in machinery components (e.g., gears and bearings) is an important category of such diagnostic information that is our focus herein. Almost all physical techniques for detecting the onset of such defects, e.g., fatigue cracks or surface spalling, involve the measurement of abnormal vibration signals transmitted through structural media due to the presence of the defect. Vibration sensors (e.g., accelerometers) may be used to measure these signals.

Sensitive spectral analysis and astute signal processing of such sensor data is critical, chiefly because in practical machinery operational scenarios, there are an enormous number of vibrations that interfere with one another and result in an observed signal that appears bewildering and uninformative. The main problem is therefore one of discovering the needle (i.e., spectral content that provides useful and pertinent information about an incipient fault) in the haystack (i.e., the raw vibration signal). A number of sophisticated techniques

have been proposed specifically for treating this very problem, including bispectral analysis [14], envelope spectral analysis [8], statistical moment analysis [9], time-frequency/time-scale analysis [2, 4], signal averaging [6], and signal template-based demodulation [7]. The latter approach is the one explored herein. In particular, the approach involves decomposing the vibration signal from a complex system, such as a gearbox, into amplitude modulation and phase modulation signals. The presence of such may be justified on physical grounds since amplitude and phase modulation arise, respectively, from radial and tangential forces exerted during impact between a defective region of one surface and the surface of another contacting member [11]. To present our signal demodulation approach for gears and bearings, we first work through its application to a gear problem. The method is then extended to the more complicated problem of bearing vibrations. The basic methodology and key assumptions of the proposed signal-processing approach are discussed. The computational complexity of the method is addressed, and a means of carrying out the computational tasks using neural network methods is presented.

Joint Amplitude/Phase Demodulation of Gear Vibrations: Consider a defect-free gear. Under normal, constant-load operation, the vibration signal, $y(t)$, emitted by the gear is periodic in the gear-mesh frequency, f_{gm}, *viz.*,

$$y(t) = \sum_{k=-\infty}^{\infty} y_k e^{ikN\omega_{\mathrm{gm}}t} \tag{1a}$$

in which $\omega_{\mathrm{gm}} = 2\pi f_{\mathrm{gm}}$ and N is the number of teeth. The existence of vibrations in a defect-free gear reflects the facts that the structural materials are not infinitely rigid and that the transmission force varies slightly with the angular displacement of the teeth in contact with the load rack. It is highly advantageous, as a number of authors [6, 7] have pointed out, to work with the angular displacement, θ, of the gear rather than time, t, as the independent variable. This compensates for variations in the rotational velocity, $d\theta/dt$, of the gear and facilitates averaging the signal over many gear revolutions to improve the signal-to-noise ratio. We thus recast Eq. 1a as

$$y(\theta) = \sum_{k=-\infty}^{\infty} y_k e^{ikN\theta} \tag{1b}$$

which indicates that under normal loading conditions, a defect-free gear exhibits a vibration signal of angular period $2\pi/N$, since there are N gear-mesh events per gear revolution.

In the presence of a localized structural or metallurgical defect, a perturbed vibration signal is observed. The signal exhibits both amplitude and phase modulation and may be modeled [7] in the mathematical form

$$y(\theta) = A(\theta) \cdot \sum_{k=-\infty}^{\infty} X_k e^{ik[N\theta + B(\theta)]} \tag{2}$$

in which $A(\theta)$ and $B(\theta)$ both have period 2π. In the limiting case of an incipient fault, the perturbed signal should be only infinitesimally different from the unperturbed signal in Eq. 1b, which implies that $A(\theta) \to 1$ and $B(\theta) \to 0$ in the incipient fault limit. As a defect develops progressively over the life of a particular gear, $A(\theta)$ and $B(\theta)$ may be expected to grow slowly, but the X_k's will remain constant (i.e., the X_k's are independent of defect severity). It thus follows that for a defect-free gear,

$$y(\theta) = \sum_{k=-\infty}^{\infty} X_k e^{ikN\theta} \tag{3}$$

as in Eq. 1b. If the vibration signature of a gear element is measured early in life before the onset of degradation effects, the X_k's can be computed directly via a discrete Fourier

transform (DFT). Retained knowledge of these X_k values facilitates simultaneous amplitude and phase demodulation of the perturbed signal in Eq. 2.

As a result of a localized tooth defect, however small, the measured vibration signal, $y(\theta)$ now has angular period 2π rather than $2\pi/N$, since the defective tooth makes contact with the load once per gear revolution. Moreover, even in the case of multiple tooth defects, $y(\theta)$ has angular period 2π. The prefactor, $A(\theta)$, represents an amplitude modulation and may be expanded as a discrete-tone Fourier series, *viz.*,

$$A(\theta) = \sum_{k=-\infty}^{\infty} A_k e^{ik\theta} . \tag{4}$$

Since the amplitude and phase modulation effects need not necessarily be the same at the various tooth-mesh harmonics, $A(\theta)$ is *not* necessarily equal to the amplitude of $y(\theta)$ that one would obtain from computation of the Hilbert transform. The expression

$$v(\theta) = \sum_{k=-\infty}^{\infty} X_k e^{ik[N\theta + B(\theta)]} \tag{5}$$

that accompanies $A(\theta)$ in Eq. 2 also has angular period 2π and may be expanded as a Fourier series, *viz.*,

$$v(\theta) = \sum_{r=-\infty}^{\infty} V_r e^{ir\theta} . \tag{6}$$

Note that $v(\theta)$ is of the same phase as $y(\theta)$. This important result follows from a theorem [1, 10] that states that if $f(t)$ is a low-frequency signal and $g(t)$ is a high-frequency signal lying completely above $f(t)$ in the frequency domain with no overlap, then the Hilbert transform of the product signal decomposes as

$$H[f(t)g(t)] = f(t)\, H[g(t)] \tag{7}$$

where the operator H denotes the Hilbert transform. It follows that if $f(t)$ (corresponding to $A(\theta)$) is in the form a real-valued signal and $g(t)$ (corresponding to $v(\theta)$) is in the form of an analytic signal (*viz.*, of the form $s(t) + i\, H[s(t)]$) then the instantaneous phase of $f(t)g(t)$ is identically equal to that of $g(t)$. Note that $A(\theta)$, for an incipient defect, is a narrowband signal centered at frequency zero, whereas the terms in the expression for $v(\theta)$ (Eq. 5) are narrowband signals centered at frequency kN (in which X_0 can be set to zero without loss of generality).

Since the phase of $v(\theta)$ is equal to that of $y(\theta)$, which is an observable signal, it follows that, given the X_k's, we have sufficient information to compute $B(\theta)$ uniquely. The solution procedure begins with the phase modulation, $B(\theta)$, which, like $A(\theta)$, is of angular period 2π, *viz.*,

$$B(\theta) = \sum_{q=1}^{\infty} B_q \sin(q\theta - \beta_q) \tag{8}$$

where the B_q's are real-valued. $B(\theta)$ appears in the expression for $v(\theta)$ in Eq. 5 in a factor of the form

$$W_k(\theta) \equiv e^{ik\, B(\theta)} \tag{9}$$

which, as a periodic function of angular period 2π, may be expressed as a Fourier series

$$W_k(\theta) = \sum_{h=-\infty}^{\infty} M_h(k, \underline{B}, \underline{\beta})\, e^{ih\theta} \tag{10}$$

in which $M_h(k, \underline{B}, \underline{\beta})$ may be computed simply by taking the DFT of W_k. The Fourier coefficients depend not only on the Fourier index, h, but also the Fourier amplitudes and

phases (collectively denoted as $\underline{B}$ and $\underline{\beta}$ respectively) of $B(\theta)$ and the gear-mesh harmonic index, k. Note that in the incipient fault limit, $B(\theta) \approx 0$, from which it follows that $M_h(k, \underline{B}, \underline{\beta}) \approx \delta_{h,0}$.

Having hypothesized the functional form of $B(\theta)$, i.e., $\underline{B}$ and $\underline{\beta}$, the $M_h(k, \underline{B}, \underline{\beta})$'s can be computed, and we obtain an expression relating it and the X_k's to the V_r's, *viz.*,

$$\sum_{r=-\infty}^{\infty} V_r e^{ir\theta} = \sum_{k=-\infty}^{\infty} \sum_{h=-\infty}^{\infty} X_k M_h(k, \underline{B}, \underline{\beta})\, e^{i(kN+h)\theta}\,. \tag{11}$$

The V_r's are thus computable as

$$V_r = \sum_{k=-\infty}^{\infty} X_k M_{r-kN}(s, \underline{B}, \underline{\beta})\,. \tag{12}$$

Having computed all of the V_r's, we can thus reconstruct the partial signal $\hat{v}(\theta)$, which is analytic and therefore has a well-defined instantaneous phase, which we denote as $\hat{\phi}_v(\theta)$. The caret denotes an estimate in that we have surmised $B(\theta)$, and the results reflect the accuracy of this approximation or educated guess. To judge whether the surmised $B(\theta)$ is accurate, the estimated phase, $\hat{\phi}_v(\theta)$ is required to match the actual phase, $\phi_v(\theta)$, which can be inferred directly from the observed vibration signal, $y(\theta)$. The degree to which the instantaneous phases over the angular interval $(0, 2\pi)$ coincide reflects the accuracy of the hypothesized phase modulation signal, $B(\theta)$. The discrepancy, as a *functional* of the function $B(\theta)$, may be quantified by way of the squared error over $(0, 2\pi)$, *viz.*,

$$J[B] = \int_0^{2\pi} \left[\hat{\phi}_v(\theta) - \phi_v(\theta)\right]^2 d\theta\,. \tag{13}$$

To determine $B(\theta)$ requires a search algorithm to find optimal values for $\underline{B}$ and $\underline{\beta}$. Since $\underline{B}$ and $\underline{\beta}$ both are infinite sets, however, it is feasible to proceed by approximating $\underline{B}$ and $\underline{\beta}$ as finite Fourier series and gradually including additional harmonics in more refined approximations. We later show, following the succeeding discussion on the extension of the methodology to bearings, how this numerical procedure can be implemented using neural network algorithms to enhance computational speed while retaining excellent accuracy.

As in [7], the amplitude signal, $A(\theta)$, may then be computed at the very end of the entire procedure as $y(\theta)/v(\theta)$, once the optimal $B(\theta)$ functional form has been found and $v(\theta)$ has been computed.

Joint Amplitude/Phase Demodulation of Bearing Vibrations: In a defect-free roller bearing under constant loading conditions, the vibration signal, $y(t)$, assumes the same form as in the gear problem, *viz.*,

$$y(\theta) = \sum_{k=-\infty}^{\infty} y_k e^{ikN\theta} \tag{14}$$

where N is the number of balls and θ is the angular displacement of the cage. The perturbed signal that arises in the presence of a localized structural defect also assumes the form

$$y(\theta) = A(\theta) \cdot \sum_{k=-\infty}^{\infty} X_k e^{ik[N\theta + B(\theta)]} \tag{15}$$

as in Eq. 2. Unlike in the gear problem, however, $A(\theta)$ and $B(\theta)$ have more complicated frequency-domain properties. They are both biperiodic with angular periods $2\pi\gamma_q$ and $2\pi\gamma_d$,

where

$$\gamma_q = f_{\text{cage}}/f_{\text{load-pass}} \tag{16a}$$

$$\gamma_d = f_{\text{cage}}/f_{\text{defect}}\,. \tag{16b}$$

The biperiodic form is due to the physical circumstance that the perturbative excitation force, $e(t)$, resulting from contact between a surface defect and other members of the bearing assembly, is the product of an impact amplification function, $q(t)$, and an impact event function, $d(t)$ [8]. The latter function is a train of impulses that indicate the instants at which the defective surface region contacts other members. The spacing between the impulses is determined by the frequency, f_{defect}, characteristic of the particular type of defect. The three most common examples are

$$f_{\text{BPIR}} = \frac{1}{2}Nf_{\text{shaft}}\left(1+\frac{d_r}{d_p}\cos\alpha\right) \qquad \text{for an inner-race defect;} \tag{17a}$$

$$f_{\text{BPOR}} = \frac{1}{2}Nf_{\text{shaft}}\left(1-\frac{d_r}{d_p}\cos\alpha\right) \qquad \text{for an outer-race defect;} \tag{17b}$$

$$f_{\text{roller spin}} = \frac{1}{2}f_{\text{shaft}}\left(\frac{d_p}{d_r}\right)\left(1-\frac{d_r^2}{d_p^2}\cos^2\alpha\right) \qquad \text{for a rolling-element defect,} \tag{17c}$$

where d_r is the diameter of the rolling elements, d_p is the pitch diameter, and α is the contact angle. The cage frequency, f_{cage}, is equal to f_{BPOR}/N (assuming that the outer race is stationary and the inner race is rotating at the shaft frequency) and is analogous to the gear-mesh frequency in the gear problem.

The impact amplification function, $q(t)$, reflects the fact that an impact event involving a rolling element in the load zone is more consequential than an impact involving an unloaded rolling element. $q(t)$ is therefore periodic at the frequency $f_{\text{load-pass}}$, which is the frequency at which the defective surface passes through the load zone, *viz.,*

$$f_{\text{load-pass}} = f_{\text{shaft}} \qquad \text{for an inner-race defect;} \tag{18a}$$

$$f_{\text{load-pass}} = 0 \qquad \text{for an outer-race defect;} \tag{18b}$$

$$f_{\text{load-pass}} = f_{\text{cage}} \qquad \text{for a rolling-element defect.} \tag{18c}$$

All of the various frequencies in the above equations may be expressed as products of the cage frequency times a factor that depends solely on the bearing geometry. Hence, γ_q and γ_d are independent of variations in the shaft rotational speed.

The periodicity of $A(\theta)$ and $B(\theta)$ depends strongly on the type of defect. For an outer-race defect, $\gamma_q \to \infty$ and $\gamma_d = f_{\text{cage}}/f_{\text{BPOR}} = 1/N$. An outer-race defect is therefore equivalent dynamically to a gear defect, as treated in the preceding section. For a rolling-element defect, $\gamma_q = 1$, but $\gamma_d = f_{\text{cage}}/f_{\text{roller spin}}$. The frequency ratio $f_{\text{cage}}/f_{\text{roller spin}}$ is equal to

$$\frac{f_{\text{cage}}}{f_{\text{roller spin}}} = \left(\frac{d_r}{d_p}\right)\left(1+\frac{d_r}{d_p}\cos\alpha\right)^{-1} \tag{19}$$

which is generally an irrational number. For an inner-race defect, γ_d and γ_q are both irrational. The presence of multiple frequencies and the irrationality of one or both γ's makes the mathematics of gear and bearing problems significantly different. If there are multiple faults of different types, three or more γ's may arise.

To demonstrate the demodulation procedure for a bearing, we focus on the special case of an inner-race defect, wherein the vibration signal assumes a triperiodic Fourier series form, *viz.*

$$y(\theta) = \sum_{k=-\infty}^{\infty}\sum_{d=-\infty}^{\infty}\sum_{q=-\infty}^{\infty} Y_{k,d,q}\, e^{i[kN+\gamma_d d+\gamma_q q]\theta} \tag{20}$$

with $Y_{k,d,q} \rightarrow Y_{k,0,0}\,\delta_{d,0}\delta_{q,0}$ in the incipient fault limit. Similarly, $v(\theta)$ becomes

$$v(\theta) = \sum_{r=-\infty}^{\infty} \sum_{d=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} V_{r,d,q}\, e^{i[r+\gamma_d d+\gamma_q q]\theta}\,. \tag{21}$$

The amplitude function becomes

$$A(\theta) = \sum_{d,q} A_{d,q}\, e^{i(\gamma_d d+\gamma_q q)\theta} \tag{22}$$

as in Eq. 4, and the phase modulation function becomes

$$B(\theta) = \sum_{d=1}^{\infty} \sum_{q=1}^{\infty} B_{q,d} \sin\left[(\gamma_d d + \gamma_q q)\theta - \beta_{q,d}\right]\,. \tag{23}$$

$W_k(\theta)$ becomes

$$W_k(\theta) = \sum_{h_d=-\infty}^{\infty} \sum_{h_q=-\infty}^{\infty} M_{h_d,h_q}(k, \underline{B}, \underline{\beta})\, e^{i(\gamma_d h_d+\gamma_q h_q)\theta} \tag{24}$$

with

$$M_{h_d,h_q}(k, \underline{B}, \underline{\beta}) = \delta_{h_d,0}\delta_{h_q,0} \tag{25}$$

in the incipient fault limit. The $M_{h_d,h_q}(k, \underline{B}, \underline{\beta})$'s can be computed via a DFT, as in the gear case. However, since $W_k(\theta)$ is no longer periodic over the interval $(0, 2\pi)$, the DFT is somewhat less clean numerically. This means that either a very large Fourier window must be used or that $M_{h_d,h_q}(k, \underline{B}, \underline{\beta})$ may have to be computed analytically by appealing to a Bessel function expansion of $W_k(\theta)$. This is done by substituting Eq. 23 into Eq. 9, which yields

$$W_k(\theta) = \prod_{d=1}^{\infty} \prod_{q=1}^{\infty} e^{ikB_{d,q} \sin[(\gamma_d d+\gamma_q q)\theta-\beta_{q,d}]} \tag{26a}$$

$$= \prod_{d=1}^{\infty} \prod_{q=1}^{\infty} \sum_{s=-\infty}^{\infty} J_s(kB_{d,q})\, e^{-ikB_{d,q} \sin \beta_{d,q}} \cdot e^{is(\gamma_d d+\gamma_q q)\theta} \tag{26b}$$

in which the mathematical identity

$$e^{ix \sin \theta} = \sum_{s=-\infty}^{\infty} J_s(x)\, e^{is\theta} \tag{27}$$

was invoked. It follows from inspection of Eq. 26 that the Fourier coefficients are

$$M_{h_d,h_q}(k, \underline{B}, \underline{\beta}) = \sum_{s} J_s(kB_{d,q})\, e^{-ikB_{d,q} \sin \beta_{d,q}}\, \delta_{sd,h_d}\delta_{sq,h_q} \tag{28}$$

in which the summation is over all common factors, s, that divide both h_d and h_q.

Having computed $M_{h_d,h_q}(k, \underline{B}, \underline{\beta})$, the X_k's can be computed by comparing the left- and right-hand sides of

$$\sum_{r,d,q} V_{r,d,q}\, e^{i[r+\gamma_d d+\gamma_q q]\theta} = \sum_{k,d,q} X_k M_{h_d,h_q}(k, \underline{B}, \underline{\beta})\, e^{i(kN+\gamma_d h_d+\gamma_q h_q)\theta}\,. \tag{29}$$

Since the γ's are irrational, it follows that the only combinations of indices that enable the exponential factors to agree are in pairs of corresponding terms on the two sides of Eq. 29, such that $r = kN$, $d = h_d$, and $q = h_q$. It follows that the X_k's can be computed readily from the relation

$$V_{k,d,q} = X_k M_{d,q}(k, \underline{B}, \underline{\beta}) \tag{30}$$

which, unlike in the gear case, becomes exact owing to the irrationality of the γ's. This result actually makes the irrationality of γ helpful, not a hinderance, since it simplifies the computational burden. For the gear problem, this simplification is also valid if N is large, in which case N is "almost irrational" in the sense that if it is written as p/q, where p and q are integers, p or q must be large.

The instantaneous phases of $y(\theta)$ and $v(\theta)$, as in the gear application, are required to be equal, and a search algorithm must be used, as before. Once an optimal functional form for $B(\theta)$ is found, the amplitude, $A(\theta)$, is computed at the very end.

Amplitude/Phase Demodulation Neural Network: Numerical solution of the vibration amplitude and phase demodulation problem can be realized by means of neural networks, as described herein. *GNOSIS* (*Generalized Networks for Optimal Synthesis of Information Systems*) [5, 12, 13] is a software package developed by Barron Associates, Inc., for synthesis and implementation of a wide variety of artificial neural network paradigms that employ various alternative types of basis functions. Specifically, a *GNOSIS* network can be designed to automate the algorithmic computation procedure of the flowchart in Fig. 1, which involves a search algorithm.

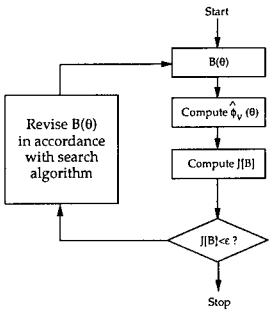


Figure 1: Search Algorithm Procedure Realized Via *GNOSIS*

The *GNOSIS* network estimates the function $\hat{\phi}_v$ over the angular range $0 \leq \theta < 2\pi$ by optimizing the phase modulation, $B(\theta)$, in such a way that the objective function

$$J[B] = \int_0^{2\pi} \left[\hat{\phi}_v(\theta) - \phi_v(\theta)\right]^2 d\theta \tag{31}$$

is minimized globally. Several numerical simulation tests were performed to construct *GNOSIS* algorithms to implement the flowchart algorithm in Fig. 1.

Several synthetic vibration signals of the form

$$y(\theta) \;=\; A(\theta) \cdot \mathrm{Re}\left\{\sum_{k=1}^{K} X_k e^{ik[N\theta + B(\theta)]}\right\}$$

$$= A(\theta) \cdot \sum_{k=1}^{K} \{(\mathrm{Re}\, X_k) \cos(k[N\theta + B(\theta)]) - (\mathrm{Im}\, X_k) \sin(k[N\theta + B(\theta)])\} \quad (32)$$

were generated for a gear with $N = 10$ teeth. K denotes the maximum number of gear-mesh harmonic terms included in the simulated signal. Synthetic $A(\theta)$ and $B(\theta)$ functions were computed as

$$A(\theta) = \sum_{p=0}^{P} [(\mathrm{Re}\, A_p) \cos p\theta - (\mathrm{Im}\, A_p) \sin p\theta] \quad (33a)$$

$$B(\theta) = \sum_{q=1}^{Q} [(\mathrm{Re}\, B_q) \cos q\theta - (\mathrm{Im}\, B_q) \sin q\theta] \quad (33b)$$

where P and Q denote the maximum number of terms in $A(\theta)$ and $B(\theta)$ respectively.

The analytic continuation, $y_a(\theta)$, of the observed signal, $y(\theta)$ was computed via the Hilbert transform, *viz.*

$$y_a(\theta) = y(\theta) + i\, H[y(\theta)] = |y(\theta)|\, e^{i\phi_y(\theta)}. \quad (34)$$

However, instead of computing the $\phi_y(\theta)$ directly, the *GNOSIS* networks were designed to model the real and imaginary parts of the normalized signal, *viz.*,

$$\cos \phi_y(\theta) = y(\theta)/|y(\theta)| \quad (35a)$$

$$\sin \phi_y(\theta) = H[y(\theta)]/|y(\theta)|. \quad (35b)$$

The normalized signals, unlike $\phi_y(\theta)$ itself, are continuous functions of θ, which facilitates the *GNOSIS* computations.

A feedforward network structure suitable for modeling the normalized signals can be realized using polynomial basis functions with trigonometric and linear post-transformations. A block diagram of the basic network structure is depicted in Fig. 2.

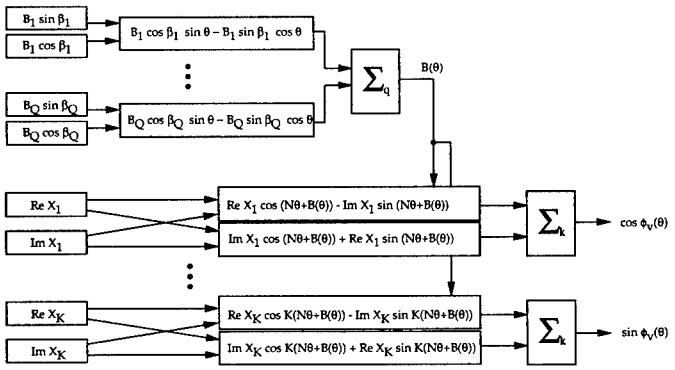


Figure 2: *GNOSIS* Network for Vibration Demodulation

The neural network algorithm for modeling the vibration signal assumed the *structural form* of a phase-modulated $v(\theta)$ signal; the Levenberg-Marquardt algorithm [12, 13] within *GNOSIS* was used to fit coefficients to the training signal. This approach is very efficient computationally. The versatility of *GNOSIS*, with respect to implementing alternative

types of basis functions and structuring networks in innovative ways pursuant to the needs of specialized applications, provides a strong incentive for adopting the neural network paradigm for the present application.

A network was trained on the signal $y(\theta)$ with the angular interval (0.2π) discretized into $2^{11} = 2,048$ sample points. Between ten and twenty iterations were required for convergence of the computed $B(\theta)$. Fig. 3 provides a comparison of the actual and *GNOSIS*-computed $B(\theta)$ signals for a particular sample problem with $K = 6$, $P = 4$ and $Q = 7$. In the figure, the computed $B(\theta)$ is sufficiently accurate that it is visually indistinguishable from the actual signal at the scale shown.

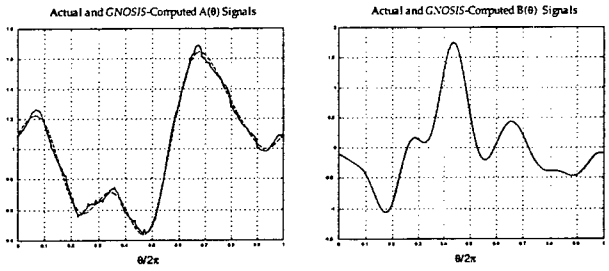


Figure 3: $A(\theta)$ (left) and $B(\theta)$ (right) Signals

Note that the X_k's are included in the network structure simply as additional degrees of freedom independent of the parameters defining $B(\theta)$. Very significantly, it was found that no *a priori* knowledge of their values was necessary. This implies that diagnostic knowledge from the early life of the gear element, as previously discussed, is actually *not* necessary after all!

A second major assumption that was made in the introductory discussion was that the $A(\theta)$ and $B(\theta)$ signals must both be sufficiently narrowband to the extent that the conditions of Eq. 7 are satisfied. Specifically, this requires that the bandwidths satisfy the inequality

$$\mathcal{B}_A + (\delta B + 1)\mathcal{B}_B < f_{\text{gm}} \tag{36}$$

where $\mathcal{B}$ denotes bandwidth and δB is the peak absolute value of $B(\theta)$, i.e., the maximum phase deviation introduced by $B(\theta)$. Eq. 36 follows from Carson's rule [3] and the fact that $v_1(\theta)$ contains the lowest frequency components of $v(\theta)$. It is important to note that in the case of $B(\theta)$, Eq. 36 not only imposes a bandwidth restriction, but also a magnitude restriction.

Fig. 4 displays the spectrum of $v(\theta)$ for the particular test case in Fig. 3. The spectrum contains significant modulation spreading into the baseband region occupied by $A(\theta)$, in violation of the conditions of Eq. 36. However, as exemplified by the good accuracy in Fig. 3, the assumption of narrowband $A(\theta)$ and $B(\theta)$ can also be lifted! Once $v(\theta)$ is computed, $A(\theta)$ can be obtained from $A(\theta) = y(\theta)/v(\theta)$. The left plot in Fig. 3 displays the actual (solid) and computed (dashed) $A(\theta)$ signals; the visible discrepancy is attributable to the failure of Eq. 7 to hold in the present case. The phase of the $v(\theta)$ signal computed in this first pass, however, can be used to apply the algorithm in a second pass to obtain incrementally better $B(\theta)$, X_k's, and $A(\theta)$. The process can be repeated until a self-consistent $A(\theta)$ is obtained. Alternatively, the computed $A(\theta)$ can be low-pass filtered to obtain a better approximation.

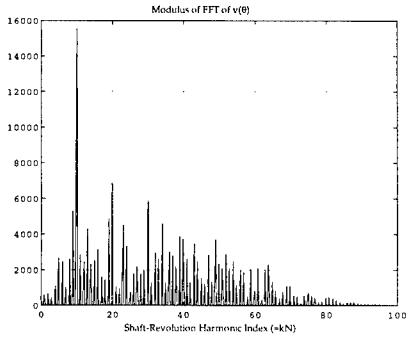


Figure 4: Fourier Spectrum of $y(\theta)$ Signal

Conclusion: We have presented a diagnostic neural network algorithm for decomposing gear and bearing vibration signals into amplitude and phase modulation components. Surprisingly few *a priori* assumptions are necessary with the approach.

References:

[1] Bedrosian, E., "A product theorem for Hilbert transforms," *Proc. of the IEEE*, Vol. 51, 1963, pp. 868-869.

[2] Brotherton, T., T. Pollard, and D. Jones, "Applications of time-frequency and time-scale representations to fault detection and classification," *Proc. IEEE-SP Int'l. Symp. on Time-Frequency and Time-Scale Analysis*, Oct. 4-6, 1992, Victoria, BC, Canada.

[3] Couch, II, Leon W., *Digital and Analog Communication Systems*, 4th ed., 1993, p. 332. (MacMillan: New York)

[4] Forrester, B.D., "Use of the Wigner-Ville distribution in helicopter fault detection," *ASSPA '89, Signal Processing, Theories, Implementations, and Applications* (R.F. Barrett, ed.), Adelaide, Australia, Apr. 17-19, 1989, pp. 78-82.

[5] Barron Associates, Inc., *Generalized Networks for Optimal Synthesis of Information Systems*, Users' Manual, Version 2.1, Oct. 1996.

[6] Jong, J., J. McBride, J. Jones, T. Fiorucci, and T. Zoladz, "Synchronous phase averaging method for machinery diagnostics," *Integrated Monitoring, Diagnostics, and Failure Prevention - Proc. Joint Conf. 50th Mtg. of the MFPT*, Mobile, AL, Apr. 22-26, 1996, pp. 441-451. (Vibration Institute: Willowbrook, IL)

[7] Ma, J. and C.J. Li, "A new approach to gear demodulation and its application to defect detection," *Advanced Materials and Process Technology for Mechanical Failure Prevention - Proc. 48th Mtg. of the MFPT*, Wakefield, MA, Apr. 19-21, 1994, pp. 43-55. (Vibration Institute: Willowbrook, IL)

[8] McFadden, P.D. and J.D. Smith, "Model for the vibration produced by a single point defect in a rolling element bearing," *J. of Sound and Vibration*, Vol. 96, No. 1, 1984, pp. 69-82.

[9] McFadden, P.D. and J.D. Smith, "A signal processing technique for detecting local defects in a gear from the signal average of the vibration," *Proc. Inst. Mech. Engrs.*, Vol. 199, No. C4, 1985, pp. 287-292.

[10] Rihaczek, A.W., "Hilbert transforms and the complex representation of real signals," *Proc. of the IEEE*, Vol. 54, 1966, pp. 434-435.

[11] Swanson, D.C., private communication.

[12] Ward, D.G., B.E. Parker, Jr., and R.L. Barron, *Principles of Function Estimation Using Artificial Neural Networks*, Barron Associates, Inc. Technical Memorandum, Sept. 1993.

[13] Ward, D.G., "Generalized networks for complex function modeling," *Proc. IEEE Systems, Man & Cybernetics (SMC-94) Conf.*, Oct. 2-5, 1994.

[14] Zoladz, T., E. Earhart, and T. Fiorucci, *Bearing Defect Signature Analysis Using Advanced Nonlinear Signal Analysis in a Controlled Environment*, CDDF Final Report, Project No. 93-10, NASA Technical Memorandum 108491, May 1995.

Acknowledgments: This investigation was funded by the Office of Naval Research under Contract N00014-95-C-0413. The authors thank Dr. Thomas M. McKenna for his kind support of this work. Gratitude is also expressed to Dr. C.J. Li of the Rensselaer Polytechnic Institute and Dr. J. Ma of the New York Institute of Technology for their inspiration for, and critical review of, our work. The authors, however, assume full responsibility for any errors in the present work.

FAULT SEVERITY ESTIMATION USING A NEURAL NETWORK FAULT TRACKING APPROACH

Jose E. Lopez
CYTEL Systems Corporation
33 Boston Post Road, West, Suite 270
Marlborough, MA 01752

Kevin Oliver
Inna A. Farber Yeldham
ALPHATECH, Inc.
50 Mall Road
Burlington, MA 01803-4562

Abstract: A much desired milestone for the machinery monitoring community is the ability to move beyond static diagnostics and develop a capacity to track an evolving failure, provide estimates on the severity of the impending fault, and develop a prognostic capability. This paper will provide some insight as to how one develops such a capability using wavelet-based fault monitoring and diagnostic kernels. Essentially, the underpinnings of such a capability result from the ability to 1) develop robust wavelet-based parameters related to an underlying assessment of the operational machine model, 2) rapid extraction of wavelet-based vectors containing the desired individual parameters and 3) post-processing of the raw outputs of a globally trained, nonlinear classifier (e.g. backpropagation neural network) for tracking of the evolving failure through feature space. This paper explores this technique by applying it to a digitized vibration record of an evolving gas turbine blade tip failure obtained during cycle testing of an General Electric F100-GE-129 turbofan gas turbine.

Key Words: monitoring; rotating machinery; fault detection; gas turbine; prognosis; condition based maintenance; wavelets; neural networks; severity estimation

INTRODUCTION: Vibration monitoring, today, still remains the main foundation upon which machinery monitoring systems are built. In recent years there has been much activity in the area of applying new and emerging signal processing technologies to the problem of machinery condition assessment and fault diagnostics [1-9]. These methods have proven effective in separating out machinery states (i.e. determining, based on the vibration record alone, whether the machine is operating in a normal mode or a faulted mode). In addition these techniques have proven effective in associating the correct fault conditions with the detected fault vibration records.

A desired milestone in the machinery monitoring community is a migration toward effective, automated, on-line, condition monitoring systems which are capable of tracking an evolving failure, providing estimates on the severity of the impending fault, and developing a credible prognostic capability. Such systems will have a huge impact on machine intensive manufacturing operations. Robust systems deployed across the manufacturing domain will provide the critical margins necessary to stay globally competitive by impacting quality control, process control and machine availability. In the military domain, such systems, will aid the goals of reduced manning across a number of platforms and provide a basis for life extension through identification and timely maintenance of degraded machinery.

This paper provides some insight as to how one develops such a capability using wavelet-based fault monitoring and diagnostic kernels [4-12]. Essentially, the underpinnings of such a capability result from the ability to 1) develop robust wavelet-based parameters related to an underlying assessment of the operational machine model, 2) rapid extraction of wavelet-based vectors containing the desired individual parameters and 3) post-processing of the raw outputs of a globally trained, nonlinear classifier (e.g. backpropagation neural network) for tracking of the evolving failure through feature space. This paper explores this technique by applying it to a digitized vibration record of an evolving gas turbine blade tip failure obtained during cycle testing of an General Electric F100-GE-129 turbofan gas turbine.

WAVELET/NEURAL NETWORK BASED METHODS: This section presents the application of wavelet-based techniques coupled with neural networks to develop a fault detection and diagnostic kernel. Continuous wavelet transforms and the selection of wavelet basis functions appropriate for real-time feature extraction are discussed.

Continuous Wavelet Transforms: To develop viable fault detection and identification (FDI) schemes, a means of extracting significant discriminate features from the vibration signal plays a critical role. Harmonic analyses in the form of a Fourier transform proves problematic for several reasons. First, the transform is global in that localized events in time can affect the entire frequency spectrum. Additionally, the Fourier transform is fundamentally not applicable to real-time monitoring applications due to the mathematical formulation of the transform that operates on the entire time axis. Windowing schemes are thus required to address the real-time feature extraction requirements for capturing important events localized in time. Unfortunately, fixed windowing schemes imply fixed time-frequency resolution in the time-frequency plane. The problem this poses is the selection of a single window that provides sufficient fidelity discriminating important events in the vibration signal that are separated by large orders of magnitude along the frequency axis. This scenario is exemplified by main helicopter transmissions where important information concerning bearings can be on the order of tens to hundreds of Hertz, whereas mesh frequencies and important fundamentals associated with gearing of the engine inputs can be on the order of tens of thousands of Hertz [9].

The continuous wavelet transform (CWT) resolves the window selection problem with a "zoom-in" and "zoom-out" capability that generates a flexible time-frequency window that automatically narrows (along the time axis) at high center-frequencies and expands (along the time axis) at low center frequencies [13]. The continuous wavelet transform provides this flexible time-frequency analysis by decomposing the vibration signal over dilated and translated wavelet basis functions. A wavelet is a function with finite energy, or a member of the function space $L^2(R)$, i.e., a wavelet function satisfies:

$$\int_{-\infty}^{\infty} |\psi(x)|^2 dx < \infty \qquad (1)$$

The wavelet function has a zero average or essentially no DC component. A set of basis functions is obtained through dilation's and translations of a base wavelet and takes the form:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) \qquad (2)$$

where u is the translation parameter and s is the dilation parameter. The wavelet transform is then achieved via the inner product of the respective vibration signal, $f(t)$, with the wavelet basis function of eq. (2):

$$W_\psi f(u,s) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt \qquad (3)$$

There are an infinite number of wavelet basis functions that satisfy eq. (1) and contain no DC component. The particular analyzing wavelet basis functions used in this work were mathematically inspired from biological systems that are effective in their decomposition and detection of vibration signals. The wavelet basis functions mimic the auditory nerve neuron's impulse response. This particular wavelet family has semi-infinite support in the time domain and can be modeled using causal real-rational transfer functions. The immediate implication is the ability to develop, on the individual wavelet basis function level, real-time feature extractors that can be efficiently implemented using auto-regressive moving average techniques (ARMA).

Wavelet-Based FDI: The goal of a wavelet-based FDI system is two-fold. First, the wavelet-based feature extraction provides the important role of extracting the essential projections of the system dynamics in an efficient manner. Second, the wavelet feature set essentially reduces the dimension of the information from the input space (real-time, continuous, analog vibration signal) to a robust lower dimensional representation that simplifies the design of the adaptive neural network classification scheme. In the design phase (as indicated in Figure 1), CWT analysis is performed on the vibration data sets to identify a set of robust wavelet features (in time and frequency) that provide discrimination between normal operation and failure conditions. In addition, these wavelet features provide sufficient separation of the failure conditions in feature space for reliable identification of the fault condition using capable classification technologies such as neural networks. Once selected, these features form the components of a feature vector extracted from the vibration signal on a regular basis. The time interval between wavelet-based feature vectors extracted from the vibration signal is user and application dependent. Most of the details of these techniques are available in [4, 6-10].

WAVELET / NEURAL NETWORK BASED SEVERITY ESTIMATION: The general wavelet-based methodology developed has essentially been used to perform robust fault detection and fault identification [4-8]. The success of these wavelet-based methods applied across a number of varied platforms has been illustrated in the following references [9-12, 15-16]. An important question to address is the extendibility of these methods to the demanding problem domain that is linked to evolving mechanical failure.

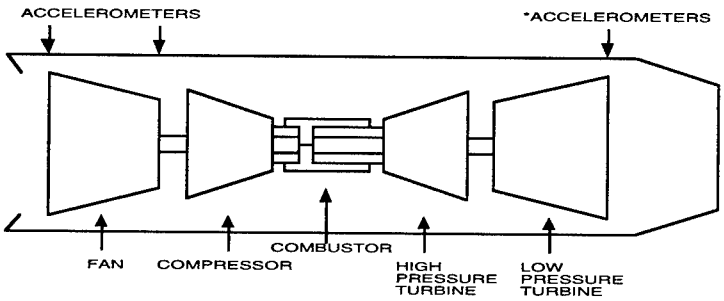
For rotating machinery, once failure has occurred, our wavelet-based methods have proven highly accurate in detecting the failure and identifying the fault. However, a desired commodity for the machinery monitoring community is the additional abilities to track an evolving failure, provide estimates on the severity of the impending fault, and develop a prognostic capability.

The following sections provide some insight as to how one develops such a capability based on our wavelet-based FDI kernels. Essentially, the underpinnings of such a capability result from the ability to 1) develop robust wavelet-based parameters related to an underlying assessment of the operational machine model, 2) rapid extraction of wavelet-based vectors containing the desired individual parameters and 3) post-processing of the raw outputs of a globally trained, nonlinear classifier (e.g.

backpropagation neural network) for tracking of the evolving failure through feature space.

EVOLVING FAILURE DATA SET DESCRIPTION: GE GAS TURBINE: In order to determine the feasibility of performing severity estimation and or prognostic algorithms an appropriate data set recording the entire record of an evolving failure is needed. Such a set was acquired from General Electric Aircraft Engine Division. The digitized data record records a blade tip failure of an F100-GE-129 turbofan gas turbine. This platform is in wide use in the aircraft industry. In addition the F100-GE-129 engine shares the same core as the LM 2500 gas turbine which is a engine deployed in the Naval surface fleet and in particular across the Aegis (CG) fleet.

Endurance tests are performed on aircraft engines to evaluate and detect component failures by operating the engine at a variety of speeds over extended time intervals. During endurance testing of the F110-GE-129 at the General Electric (GE) Aircraft Engine Division, data was collected from six accelerometers. .The accelerometers were partitioned into pairs and co-located at three distinct sites on the gas turbine frame. An accelerometer measuring horizontal vibration and an accelerometer measuring vertical vibration were positioned at each of three different sites. These three sites included the front end of the fan frame, the rear end of the fan frame, and the rear end of the low pressure turbine frame (Figure 1).



*Horizontal Accelerometer Data at this location analyzed. (Failure at Second Stage of Low Pressure Turbine.)

Figure 1. Instrumented Turbine Engine

The turbofan, unlike helicopter gas turbines, is not a fixed RPM system. Engine thrust is directly proportional to fan RPM. Essentially, the turbine fan (see Figure 1) acts as a low pressure compressor. The air is further compressed by the compressor stage. The compressed air oxidizes the engine fuel and this mixture is ignited in the combuster. The flow of hot gases across the high pressure turbine and subsequently low pressure turbine is converted into rotational motion to drive the compressor and fan. The fan effectively is a supercharged propeller, generating the necessary thrust required to propel the aircraft. An important observation, as illustrated in Figure 1, is that the fan and compressor are not mechanically linked but aerodynamically linked. The compressor / high pressure turbine core are physically linked and the fan / low pressure turbine are physically linked. The entire system is then in a concentric shaft arrangement. To increase thrust, fan speed is increased, which effectively implies increased fuel flow to the turbine combustor. The turbine blade tip shroud is a component of the turbine blade, effectively arranged in "T"

fashion relative to the blade, which ensures the efficient energy conversion by forcing the hot gas flow to occur over the turbine blades.

During the endurance tests, a shroud failure (i.e., partial separation of the shroud) occurred on a blade tip in the second stage of the low pressure turbine (LPT). This failure caused a rotor imbalance, which consequently generated engine vibrations. At the time of failure, the turbine was operating at approximately 8600 rotations per minute (rpm). GE Aircraft Division provided the raw data from all six accelerometer sites to ALPHATECH, Inc., for analysis. The purpose of the analysis was to determine whether the failure could be detected, and whether that failure could be predicted.

SELECTING AN APPROPRIATE SENSOR SITE: Analysis of the GE turbine engine data began by examining the accelerometer data from all six accelerometer locations using time-scale analysis (i.e., CWT visualizations), while considering turbine engine mechanics and general turbine engine operation. Although the time-scale analysis indicated all six sensors detected the fault, the accelerometers at the turbine fan frame location generated the most prominent reaction to the blade tip shroud failure. The attenuated response at the remote accelerometers (relative to the LPT stage 2 turbine blade) is explained by the mechanical structure of the turbine engine.

The horizontal and vertical turbine fan frame accelerometer CWT data for the period during the actual blade separation is shown in Figure 2. From the CWTs, it was evident that the horizontal accelerometer's response to the fault was more severe than the vertical accelerometer at that same location. An analysis of the pre-failure and post-failure horizontal turbine fan accelerometer data ensued.

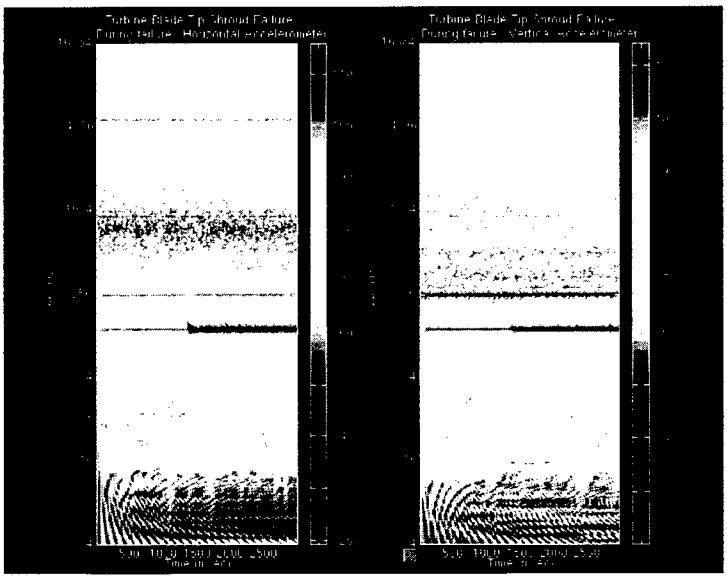


Figure 2. Turbine Fan Accelerometers During Failure:
Horizontal (left) and Vertical (right)

SEVERITY ESTIMATION: GENERAL APPROACH: Using data from the horizontal accelerometer at the turbine frame, time-scale analysis was performed on a four second region, eighty seconds prior to the shroud failure, and a four second period, one-hundred forty-one seconds following the shroud failure. By observing the Normal (i.e., pre-failure) operation CWT and comparing with the Fault operation (i.e., post-failure) CWT, several distinguishing characteristics appear (Figure 3). It was thus clear from the CWT visualizations that fault detection was possible.

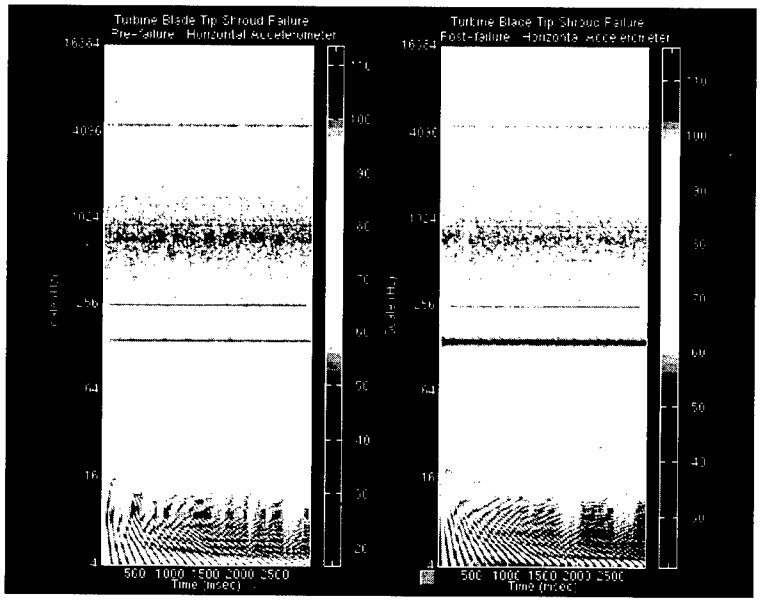


Figure 3. Turbine Fan Horizontal Accelerometer CWTs:
Before (left) and After (right) Failure

As mentioned previously, the turbine speed at the fault instant was 8600 rpm, or 143.3 Hz. Turbine speed refers to the fan RPM. Hence the line at 143.3 Hz in the CWT corresponds to the wavelet-based vibratory response obtained in reaction to the fan / LPT rotations. The very visible line at approximately 256 Hz corresponds to the wavelet-based vibratory response obtained in reaction to the core rotation (i.e. compressor / high pressure turbine combination).

The line at 143.3 Hz is noticeably affected (Figure 3) by the shroud failure, increasing in magnitude and bandwidth in the post-failure CWT. This is to be expected since this wavelet structure is due to the rotation of the physically linked fan / LPT subsystem and the failure occurred on a second stage blade tip shroud within the LPT. Other frequencies of interest, due to wavelet structural differences between the pre-failure and post-failure cases, occur at approximately 280 Hz, 500 Hz, 4550 Hz, and 6800 Hz.

Although CWT visualizations indicate that wavelet technology produces information to recognize the fault scenario, being able to reliably predict an impending failure before it actually occurs is a highly desired commodity in the turbine engine diagnostic

community. The information obtained by comparing pre and post-failure CWT time-scale images was therefore used to obtain a feature set and corresponding feature vectors that trained a global Neural Network (NN). Once the NN was satisfactorily trained (i.e., sum-squared error of .02), GE data from the failure period was applied to the neural network for processing and classification. A fifty-eight second interval was selected. The blade tip shroud failure occurred at approximately fifty-one seconds into this record.

Output from the NN occurred every 10 ms during the course of the fifty-eight seconds monitored. The raw outputs of the NN were processed by a set of causal algorithms (i.e., algorithms that could be used in an on-line, real-time monitoring system). Figure 4 presents the results of this processing.

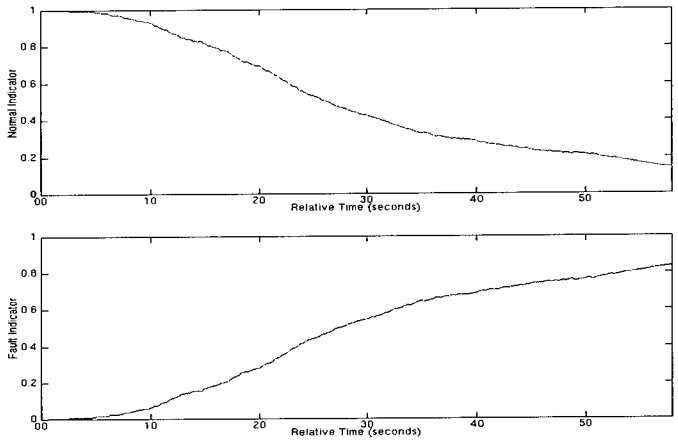


Figure 4. Fault Indicators for F110-GE-129 Turbine

PERFORMANCE RESULTS AND CONCLUSIONS: As the top plot (Normal Indicator) of Figure 4 illustrates, at the beginning of the fifty-eight second interval, or nearly 40 seconds before the actual failure, the system is generating a strong indication of normal activity, with no indication (bottom plot, Fault Indicator) of any fault; however, almost immediately thereafter, the Normal Indicator begins declining at approximately the same rate as the rise in the Fault Indicator, thereby illustrating a system with a progressively increasing fault. By approximately second 26 on the relative time plots, both Normal and Fault Indicators are at a value of 0.5, strongly implying a faulted system, particularly when compared to the Normal operation as evidenced at relative time 0-5 seconds. By second forty, or eleven seconds before actual failure, the Fault Indicator, at value 0.7, heavily outweighs the Normal indicator at less than 0.3. By second forty, the system outputs have moved closer spatially to the fault condition.

The severity estimation results are extremely promising for detecting faults before actual failure. By implementing thresholding logic upon globally trained NN data as shown in Figure 4, system operation may be accurately characterized for reliable and robust prognostic systems.

Information obtainable based on this preliminary investigation, appears to be 1) robust tracking of an evolving mechanical failure 2) the ability to assign fault severity estimates

on the basis of the progression of this wavelet-based fault tracking and 3) a prognostic capability for incipient failures linked to the underlying operational model through the judicious selection of associated wavelet-based parameters.

REFERENCES:

1. Yan, Tinghu, Zhong, "Artificial Neural Network Technique and Its Applications to Rotating Machinery Fault Diagnosis", *J. of Vib. Engrg.*, Vol. 6, pp. 205-212, 1993.
2. Tansel, I.N., Mekdeci, C., and McLaughlin, C., "Detection of Tool Failure in End Milling With Wavelet Transformations and Neural Networks", *Manufacturing Science and Engineering ASME, Production Engineering Division*, Vol. 64, pp. 369-374, 1993.
3. Rohrbaugh, R.A., "Application of Time-Frequency Analysis to Machinery Condition Assessment", *Proc. 27th Asilomar Conf. on Sigs., Syst.* , Vol. 2, pp. 1455-1458, 1993.
4. Lopez, J.E., R.R. Tenney, and J.C. Deckert, "Wavelet Feature Extraction For Real-Time Neural Network Condition Based Maintenance," *Proc. US Navy Conference on Neural Network Applications*, Arlington, VA, 16-17 June 1994, pp. 109-121.
5. Lopez, J.E., Deckert, J.C., Tenney, R.R., "Condition-Based Machinery Maintenance: Interim Report", Alphatech, TR-663, August 1994.
6. Lopez, J.E., R.R. Tenney, and J.C. Deckert, "FDI Using Real-Time Wavelet Feature Extraction," *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, 25-28 October 1994, pp. 217-220.
7. Lopez, J.E., R.R. Tenney, and J.C. Deckert, "Improved Fault Identification Using Multisensor Wavelet-Based Differential Features," *Annual Symposium on Machinery Failure Prevention Technology*, 18-20 April 1995, Virginia Beach, VA.
8. Lopez, J.E., Mc Kenney, A., "Next Generation Testing and Machine Monitoring Systems Based on Application of Wavelet and Neural Network Technologies," *Test Technology Symposium XXI: Testing in lhe 21st Century*, Laurel, MD, pp. 11-13 April, 1995.
9. Lopez, J.E., Polyak, N., "Wavelet-Based Diagnostics for Helicopter Main Transmission", *American Helicopter Society 51st Annual Forum*, Fort Worth, TX, 9-11 May 1995.
10. Lopez, J.E.,"Performance of Wavelet / Neural Network Fault Detection Under Varying Operating Points", *Proceedings of the 66th Shock and Vibration Symposium,* Vol. 1, pp. 209-217, Oct.30 - Nov. 3, Biloxi, MS 1995.
11. Lopez, J.E., Oliver, K., "Improved Analysis Tools for Wavelet-Based Fault Detection", *IASTED International Conference, Signal and Image Processing - SIP-95*, Las Vegas, NV, November 20-23, 1995.
12. Lopez, J.E., Yeldham, I.F., Oliver, K., Protz, M., "Hierarchical Neural Networks for Improved Fault Detection Using Multiple Sensors", *The American Helicopter Society 52nd Annual Forum*, pp. 1752-1758Washington, D.C., June 1996.
13. Grossman, A., R. Kronland-Martinet, and J. Morlet, "Reading and Understanding Continuous Wavelet Transforms," in *Wavelets, Time-Frequency Methods and Phase Space*, J.Combes, et. al. (Eds.), Springer-Verlag, 1989.
14. Smith, J.D., *Gears and Their Vibration*, MacMillan Press Ltd.,1983.
15. Lopez, J.E., Yeldham, I.F., Oliver, K., "Overview of Wavelet/Neural Network Fault Diagnostic Methods Applied to Rotating Machinery", *The Integrated Monitoring, Diagnostics and Failure Prevention Technology Showcase Conference,* pp. 405-417, Mobile, AL, April 1996,.
16. Lopez, J.E., Oliver, K., Yeldham, I.F., "Noninvasive Monitoring using Acoustic Sensors Coupled with Wavelet / Neural Network Based Processing", *Proceedings of the 67h Shock and Vibration Symposium,* Vol. 1, pp. 101-112, Nov. 18-22, Monterey, CA 1996.

ACKNOWLEDGMENTS: This work was supported by the Office of Naval Research under contract N00014-93-C-0077.

ON-LINE WEAR ASSESSMENT OF MILLING INSERTS USING TIME-FREQUENCY DISTRIBUTION AND NONLINEAR VIRTUAL SENSOR

C. James Li
Department of Mechanical Engineering,
Aeronautical Engineering and Mechanics
Rensselaer Polytechnic Institute, Troy, New York 12180

Tzong-Chyi Tzeng
Department of Mechanical Engineering,
Aeronautical Engineering and Mechanics
Rensselaer Polytechnic Institute, Troy, New York 12180

Abstract: This study established a way to determine the extent of milling insert wear using translational vibration data measured from a face milling process. It was shown that torsional vibration, though more difficult to measure, is more useful than translational vibration in determining insert wear. Using a nonlinear modeling technique, the dynamical relationship between the translational vibration and the torsional vibration is extracted. This relationship is then used to establish a virtual sensor that estimates hard-to-measure torsional vibration from easily measured translation vibration. Subsequently, time-frequency analysis and neural networks are used to estimate the extent of flank wear in a milling insert under single and multi-insert configurations. Experimental results have shown the usefulness of the proposed solution under various cutting conditions.

Key Words: Choi-Williams time-frequency distribution; condition monitoring; virtual sensor; wear assessment

INTRODUCTION: Undetected tool breakage may result in extensive damage to workpieces and machinery, while downtime associated with unpredicted tool breakage can be expensive, in terms of time and financial cost. Also, real-time knowledge of the state of tool wear will enable optimal scheduling of tool changes. Several methods have been developed to monitor cutting tool conditions. These methods are categorized into direct and indirect methods [1]. Direct methods include optical [2], wear particle [3], radioactivity [4], tool/workpiece junction electrical resistance [5], pneumatic [6] and tool/workpiece distance measurements [7], [8]. Indirect methods include measuring cutting force [9 -14], acoustic emission [15 - 17], power/motor current [18], vibration [19], cutting temperature [20], and roughness [21], either alone or in combination [22]. These studies deal mostly with turning operations which use non-rotating single point tools.

In the case of milling operations, unique difficulties arise as the cutting edges of the multipoint tool perform interrupted cutting on varying cutting geometries. More than one

cutting edge may be active at the same time. Moreover, the chip thickness varies throughout the cut for each cutting edge. In addition, the rotating motion of milling cutters makes the implementation of sensors close to the cutter-workpiece interface difficult.

Due to the aforementioned reasons, previous efforts in milling tool condition monitoring have mostly avoided the rotating tool problem by using a dynamometer between the workpiece and table to measure cutting forces or using an AE transducer attached to the workpiece to measure acoustic emission. However, a dynamometer reduces overall rigidity and is not flexible enough to accommodate large and heavy workpieces. An acoustic emission sensor can get in the way of fixturing and loading/unloading the workpiece, and AE's signal generating mechanism as well as its correlation to the state of the tool are not well understood.

Due to the ease of use and maturity of sensor technology, translational vibration has also been used, and its degree of success often depends on the proximity of the transducer to the tool-workpiece interface. The rotating tool in the milling process prevents the placement of the transducer very close to the interface and limits its usefulness.

This study first confirms the superiority of a different kind of vibration, i.e., torsional vibration. However, torsional vibration is not as easy to measure as translational vibration. Based on nonlinear modeling, a virtual sensor which converts translational vibration into torsional vibration is established. A diagnostic algorithm based on time frequency and neural network methods is devised to determine insert wear from the virtual sensor output. Finally, experimental validations with real data are performed.

EXPERIMENTAL SETUP: Figure 1 depicts our experimental setup, which consists of a Bridgeport vertical milling machine with a 3-inch diameter multi-insert face milling cutter installed. On the top of the machine, there is a torsional accelerometer consisting of a 4-inch diameter, 0.08 inch thick aluminum disk which rotates with the free-end of the spindle and a stationary Hoodwind instruments, series 34E acceleration sensor head. A translational vibration sensor, i.e., a PCB 307A accelerometer, is on the spindle housing. The mild steel workpiece is 2 inches wide and 6 inches long and is placed with its longitudinal axis aligned with the direction of feed. The standard feedrate, depth of cut and spindle speed were chosen to be 0.1 mm/rev, 1 mm and 325 rpm, respectively. TPG-C5 carbide inserts were installed.

Data acquisition was performed with a 486 based computer and a Metrabyte DAS-20 multi-channel data acquisition board. The sampling frequency was 25 kHz using an anti-aliasing filter with a 10 kHz cut-off frequency. Channel one digitized the translational vibration signal which comes from the translational accelerometer whose sensitive direction is in the direction of feed. Channel two digitized the torsional vibration transducer's output. Figure 2 shows samples of the two signals respectively. Each signal contains two rotations worth of data.

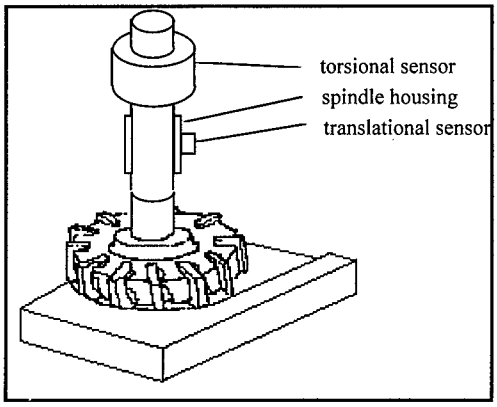


Figure 1. The experiment setup

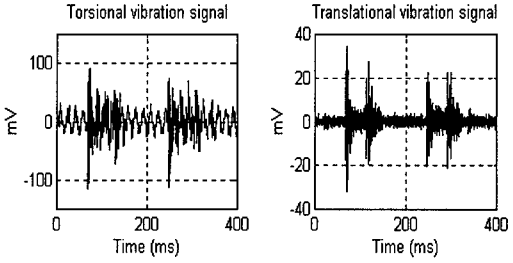


Figure 2. Typical torsional and translational vibration signals

In order to determine the effect of cutting conditions on the amplitude of the measurements (in terms of standard deviation) new inserts were used to make a number of passes under different cutting conditions including cutting speeds at 135, 210, and 660 rpm, feedrates at 0.049 mm/insert, 0.1 mm/ insert, 0.154 mm/ insert, 0.24 mm/ insert and depths of cut at 1.5, 1.2, 1.0, 0.8, 0.6 and 0.4 mm. It was found that the amplitude of the torsional vibration measurements is approximately proportional to (depth of cut, mm $)^{0.3}$, (the spindle speed, rpm$)^{0.3}$ and (feed rate, mm/insert$)^{0.1}$. Naturally one can normalize different measurements with respect to the aforementioned standard cutting conditions.

The insert wear tests were conducted in the following way. In each test, the milling cutter, holding either a single insert or 5 evenly spaced inserts, makes 12 passes under the aforementioned cutting conditions to remove material from the top of the workpiece. Measurements from both vibration sensors are taken in each of the twelve passes. Each measurement is approximately five revolutions long. The width of the flank wear is measured after each pass. In the single insert configuration, 3 tests were conducted. In the multi-insert configuration, 4 tests were conducted.

SIGNAL PROCESSING: Data from the single insert tests were analyzed first. After removing the 60Hz line frequency component from the torsional vibration data, each measurement is then segmented into 5 pieces corresponding to 1 spindle rotation each. Slight variation in rotational speed causes the segments to contain different numbers of samples. Linear interpolation is used to make every segment have the same number of points. A Fourier transform is carried out for each segment. Finally, the Fourier transforms are averaged and the result is converted back to the time domain.

The Choi-Williams time-frequency distribution (CWD) [23, 24] is calculated for the averages. CWD can be thought as a two-dimensional matrix where time and frequency are the two indices. Figure 3 shows the CWD's of a new insert and a worn insert, in the form of contour plots. The CWD's cover a time period between 0 and 11 msec. (where initial contact between the insert and the workpiece occurs at 2 msec.) and frequency from 0 to 9.375 kHz, which corresponds to a 250 by 75 matrix. Comparing the two plots, the worn insert has more energy at frequencies above 3 kHz between 4 and 8 msec., i.e., 2-6 msec. after the initial contact. This difference is still visible at a somewhat lower resolution in time and frequency. A lower resolution reduces the number of CWD elements that have to be examined by the wear size estimator in the next stage of signal processing. After systematic trials, it was found that a 10 and 15 reduction in time and frequency resolutions, respectively, has little effect on the final accuracy of the wear size estimate. Resolution is accordingly reduced by merging all the elements in a 10 by 15 area into a single element whose value is the sum of all the 150 elements. This results a reduced resolution CWD represented by a 25 by 5 matrix.

To identify the set of CWD elements that provides the most discriminant information among all the possible sets containing the same number of elements, scattering matrices [25] and orthogonalization are used. Scattering matrices are applied to identify the best element among the 125 elements of the reduced resolution CWD. Orthogonalization is then used to make the remaining elements orthogonal to the chosen one. Then scattering matrices are used again to identify the best among the remaining elements. This process chose 15 elements to be our most useful features. In Fig. 4, the time-frequency area covered by these 15 elements are marked by "x".

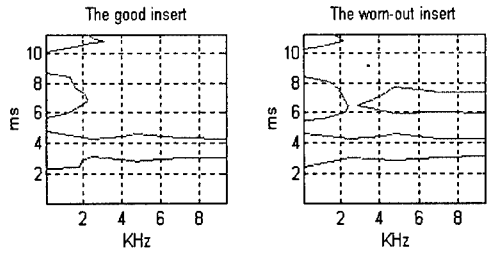


Figure 3. The CWD plots of the new and worn-out insert.

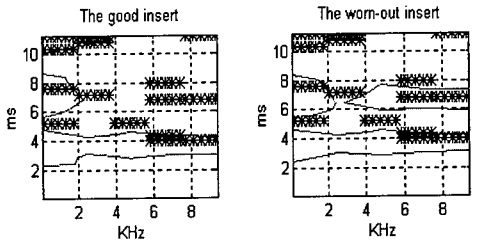


Figure 4. The fifteen most critical time-frequency areas in the CWD plot.

In the following, a neural net will be constructed to estimate the size of wear from the values of the 15 CWD elements. Among the 36 sets of data collected from 3 tests which make 12 passes each, data sets from the first and sixth passes are saved as testing data. Therefore, one has 30 data sets available for training the neural net. Furthermore, another 210 training data sets are artificially generated by adding 10% random noise to the 30 real data sets. This is done to ensure the trained network's ability to generalize. Using the Augmentation by Training with Residuals algorithm [26], a feedforward neural network that consists of 15 input, 6 hidden, and 1 output neurons is found to be able to provide a flank wear size estimate with a maximal error of 0.08mm. Figure 5 shows the estimate provided by the neural network versus the true wear size in a scattering plot where a diagonal line represents the optimal estimate. The two "*" in the figure correspond to the testing data corresponding to the first and sixth passes, obtained under the same cutting conditions as those of the training data. Their estimation error is smaller than 0.02 mm. An additional test data set which was acquired under the same cutting conditions except with a smaller depth of cut, i.e., 0.6 mm, is shown by the "×" which has an estimation error of 0.07mm.

In the previous sections, it has been clearly shown that the combination of torsional acceleration measurement, the Choi-Williams distribution, and a neural network is effective in estimating flank wear of a milling insert. However, the benefit of torsional vibration does not come without a price. The moderately priced but bulky sensor requires that a free-end of the spindle axis be fitted with a thin aluminum disk which could interfere with tool changes and be susceptible to damage. It would be desirable to use a simpler and cheaper sensor, such as an accelerometer, which is smaller, lightweight, easier to install, without sacrificing performance significantly. Unfortunately, it was found that the same signal processing strategy does not yield a satisfactory result when translation vibration is used directly. Furthermore, an array of other techniques including, for instance, spectrum analysis, McFadden's demodulation [27], and energy operator demodulation [28, 29] were similarly unsuccessful in extracting any obvious trend correlating to the size of the flank wear from translation vibration either.

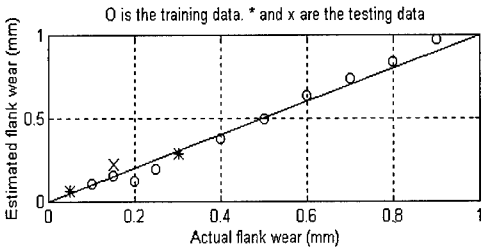


Figure 5. Estimated wear vs. actual wear

This study develops a virtual sensor by using nonlinear dynamic modeling to identify the relationship between the translation and torsional vibrations from simultaneous measurements of both. Once this relationship is available in the absence of the torsional accelerometer an approximation to the torsional acceleration can be constructed by knowing the translation vibration alone.

Since Fig. 3 shows that the torsional vibration is the most different during the period between 4 and 8 ms, the dynamic relationship between the two measurements will be identified by only using measurements around this period. Specifically, only 50 samples before and 250 samples after the initial contact are used. One thousand eight hundred samples from 6 out of the 12 passes made by one of the inserts were used to identify the model, and the rest were used as testing data. The model used is a 2nd degree nonlinear output-error model whose order and delay are determined to be 38 and 2, respectively, using the optimal linear output-error model [30].

A 38 order, 2nd degree nonlinear output-error model could contain, as many as, 39 linear terms and 741 nonlinear terms. This is obviously excessive under most conditions. A Genetic Algorithm based automated structure learning algorithm [31] was employed to identify the significant terms and their appropriate coefficients. The algorithm results in a model consisting of 250 terms and, when tested with a set of testing data, it generates the time series and time-frequency distribution shown in Figures 6 and 7. As one can see from the figures, the model is able to provide an observation of the torsional vibration that is not very different from the actual one measured by the torsional accelerometer. Its fidelity with large magnitude peaks is particularly important to this application.

Torsional vibrations are produced by the virtual sensor from the measurements of translation vibration data. The same signal processing algorithm, including the CWD and the neural network, is then used to process all these virtual measurements to estimate the size of flank wear. Figure 8 summarizes the results obtained using the virtual sensor. Those circles in the figure represent the training results, which has a maximum estimation error of 0.05 mm. The two "*" in the figure correspond to the testing data obtained under the same cutting conditions as those of the training data. Their deviation from the diagonal line is in line with the training errors. The "x" corresponds to an additional testing measurement taken with a smaller depth of cut, i.e., 0.6 mm. It has an estimation

error of 0.089 mm. The use of virtual measurements does not affect the accuracy of wear estimation significantly. The use of virtual sensor has enabled the assessment of flank wear from the translational vibration alone and therefore the sensor is no longer necessary except during the virtual sensor calibration phase.

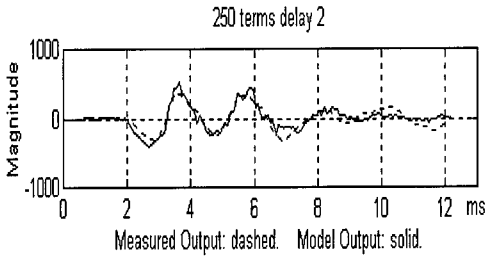


Figure 6. Virtual sensor output compared to the actual measurement

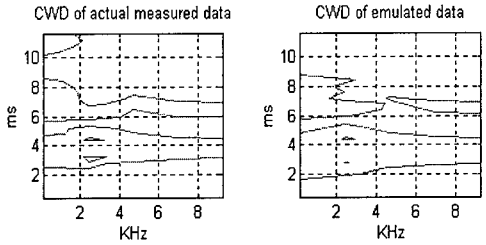


Figure 7. CWD of the time series in Fig. 6.

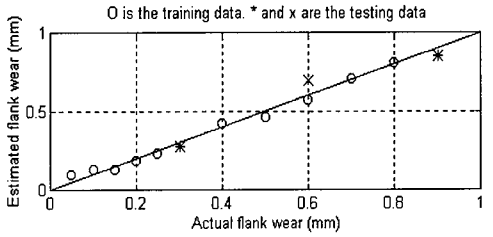


Figure 8. Estimated wear vs. actual wear based on the virtual sensor

The aforementioned methods are tested with data collected from the multi-insert tests. In this case, more than one insert is in contact with the workpiece during the critical period from 2 to 8 ms after the initial contact between the insert and the workpiece. Thus the signal in this period contains additional noise from the other insert. This extra noise could make diagnosis difficult. The results are shown in Fig. 9. The training error is

0.011 mm and testing error is 0.04 mm. It is obvious that the wear has been assessed with reasonable accuracy.

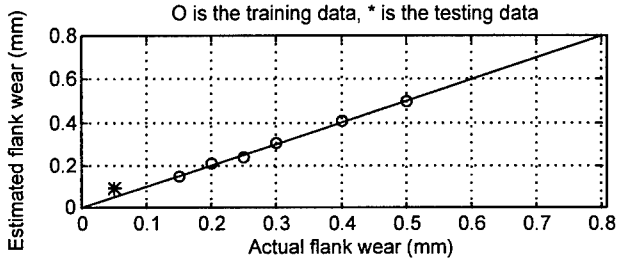


Fig. 9 Estimated flank wear vs. actual flank wear in mm (1 mm depth of cut, 325rpm and .1mm/insert feedrate with 5 inserts installed)

CONCLUSIONS: The virtual sensor, based on nonlinear modeling, and the diagnostic algorithm, based on time-frequency and neural network methods, have demonstrated their ability to determine the amount of tool wear for both single and multi-insert configurations. The estimation of flank wear was found to be accurate. Possible future work could include 1) the forecasting of insert wear from its present state and usage and 2) the identification of modal parameters of the milling machine and their relationship to the dynamics between the two vibrations.

Acknowledgment: This work was supported by Dept. of Navy under N61533-94-k-0013.

REFERENCES

1. Li, D., and Mathew, J., 1990, "Tool wear and failure monitoring techniques for turning - A review", Int. J. Mach. Tools Manufact., Vol. 30, No. 4, pp. 579-593.
2. Jeon, J. U., and Kim, S. W., 1988, "Optical flank wear monitoring of cutting tools by image processing", Wear, Vol. 127, pp. 207-217.
3. Uehara, K., 1973, "New attempts for short time tool-life testing", Annls of the CIRP, Vol. 22/1, pp. 23-24
4. Merchant, M. E., 1953, "Radioactive Cutting Tools for Rapid Tool-Life Testing", Transactions of ASME, Vol. 75, pp. 549-559.
5. Williamson, A. J., 1971, "Constriction-Resistance Concept Applied to Wear Measurement of Metal Cutting Tools," Proc. of the IEEE, Vol. 18, No. 2, pp. 381-386.
6. Micheletti, G. F., Koenig, W., and Victor, H. R., 1976, "In-process Tool Wear Sensors for Cutting Operations", Annals of the CIRP, Vol. 25, No. 2, pp. 483-488.
7. Giusti, F., Santochi, M., and Tantussi, G. 1987, "On-line sensing of flank and crater wear of cutting tools", Annals of the CIRP, Vol. 36/1, pp. 41-44.
8. Suzuki, H. and Weinmann, K. J., Nov. 1985, "An on-line tool wear sensor for straight turning operations", Journal of Engineering for Industry, Vol. 107, pp. 397-399.

9. Altintas, Y., and Yellowley, I., May. 1989, "In-process detection of tool failure in milling using cutting force models", Journal of Engineering for Industry , Vol. 111, pp. 149-157.
10. Elbestawi, M. A., Papazafiriou, T. A. and Du, R. X., 1991, "In process monitoring of tool wear in milling using cutting force signature", Int. J. Mach. Tools Manufact., Vol. 31, No.1, pp. 55-73.
11. Montgomery, D. and Altintas, Y., May. 1991, "Mechanism of cutting force and surface generation in dynamic milling", Transactions of the ASME, Vol.113, pp. 160-168.
12. Tarn, J. H., and Tomizuka, M., Aug. 1989, "On-line monitoring of tool and cutting conditions in milling", Journal of Dynamic Systems, Measurement and Control, Vol. 111, pp. 206-211.
13. Teitenberg, T. M., Bayoumi, A. E. and Yucesan, G., 1992, "Tool wear modeling through an analytic mechanistic model of milling processes", Wear, Vol. 154, pp. 287-304.
14. Victor, H., 1980, "Computer-aided measurement of cutting forces applied to the wear of an end milling cutter", Wear, Vol. 62, pp. 185-195.
15. Diei, E. N., and Dornfeld, D. A., May 1987, "Acoustic emission from the face milling process-the effects of process variables", Transactions of ASME, Vol. 109, pp. 92-99.
16. Kakade, S., Vijayaraghavan, L., and Krishnamurthy, R., 1994 , "In-process tool wear and chip-form monitoring in face milling operation using acoustic emission", Journal of Materials processing Technology, Vol. 44, pp. 207-214.
17. Rice, J. A. and Wu, S. M., Feb. 1994, "Acoustic emission source and transmission path characterization through homomorphic processing " , Transactions of ASME, Vol. 116, pp. 32-41.
18. Cuppini, D., D'Errico, G., and Rutelli, G., Aug. 1990, "Tool wear monitoring based on cutting power measurement", Wear, Vol. 139, pp. 303-311.
19. Jiang, C. Y., Zhang, Y. Z, and Xu, H. J., 1987, "In-process monitoring of tool wear stage by the frequency band-energy method", Annals of the CIRP, Vol. 36/1, pp. 45-48.
20. Groover, M. P., Karpovich, R. J., and Levy, E. K, 1977, "A study of the relationship between remote thermocouple termperatures and tool wear in machining", International Journal of Prod. Research, Vol. 25, No. 2, pp. 129-141.
21. Spirgeon, D., and Slater, R. A., 1974, "In-process indication of surface roughness using a fiber-optics transducer", Proc, 15th Int. Machine Tool Design and Research Conf., pp. 339-347.
22. Rao, D. J., 1986, "Tool wear monitoring through the dynamics of stable turning", Journal of engineering for industry, Vol. 108, pp. 183-190.
23. Choi, H., and Williams, W. J., June 1989, "Improved time-frequency representation of multicomponent signals using exponential kernels", IEEE Transactions on Acoustics, Speech, and Signal processing , Vol. 37, No.6, pp. 862-871.
24. Barry, D. T., Feb 1992, "Fast calculation of Choi-Williams Time-Frequency distribution", IEEE Transactions on signal Processing ,Vol. 40, No.2, pp. 450-455.
25. Duda, R. O., and Hart, P. E., 1973, "Pattern Classification and Scene Analysis", Addison Wesley, Reading, MA.

26. Li, C. J., and Kim, T., 1995, "A New Feedforward Neural Network Structural Learning Algorithm-Augmentation by Training with Residuals", J. Dynamic Systems, Measurement, and Control, Vol. 117, No. 3, pp. 411-415.
27. McFadden, P. D., Apr. 1986, "Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration", Journal of vibration, Acoustics, stress and Reliability in design, Vol. 108, pp. 165-170.
28. Kaiser, J. F., 1990, "On a simple algorithm to calculate the energy of a signal", IEEE ICASSP VI Speech analysis session, S7.3, pp. 381-384.
29. Maragos, P., Apr.1993, "On amplitude and frequency demodulation using energy operators", IEEE transactions on signal processing, Vol.41, No.4, pp. 1532-1550.
30. Ljung, L., 1987, System identification - Theory for the user, Prentice - Hall, Englewood Cliffs, N.J.
31. Li, C. J., and Jeon, Y., 1993, "Genetic Algorithm in Identifying Nonlinear Autoregressive with Exogenous Input Models for Nonlinear Systems", Proc. of American Control Conference, San Francisco, CA, June 2-4, IEEE, Piscataway, NJ, pp. 2305-2309.

MACHINERY FAILURE DIAGNOSIS USING DISCONTIGUOUS DETECTION METHOD

Yimin Shao and Kikuo Nezu
Mechanical System Engineering, Gunma University, 1-5-1 Tenjin-Cho, Kiryu, 376, Japan

Abstract: This paper describes the running condition monitoring and failure diagnosis methodology by using the no-contact detection method. This method applies the laser displacement sensor to measuring the vibration displacement of machinery. It not only has a feature of discontiguous measurement, but also installs laser sensor easily. The detection ability is discussed under different test conditions in this paper. Experimental results have shown that the technology is effective in failure diagnosis by analyzing vibration displacement signals of laser sensor after the transfer time and resolution of laser sensors were correctly selected. The method provides a new means of failure diagnosis of machine, and it enhances traditional machinery failure diagnosis technology.

Key words: Discontiguous detection, machinery failure, diagnosis

INTRODUCTION: The most critical area necessary for the success of an industrial company in today's competitive technological environment is the quality control of the product and its manufacturing processes. Being physical devices, product components and processes are subject to failure. Hence, machinery monitoring and failure diagnosis have become increasingly important.

A number of effective methods have been developed for localized faults of bearing, for example, statistical analysis[7], time and frequency field analysis[8], adaptive noise cancelling technology[3][5], bearing noise and acoustic emission technology[10][11], fuzzy and neural networks identification methods[1][2][9][12]. However, there are still three disadvantages in the present accelerometer measurement system. First, it is contiguous. It has some problems of installation regarding old machinery, and there are some troubles of shielding and insulation of measurement system. Second, it contains noises resulting from the nearness of machines or components of machines for the accelerometer signal. Third, it is difficult to extract feature of bearing faults from the vibration signals when the bearing is running at low speed.

With the development of science and technology in laser sensor, especially owing to the emergence of laser displacement sensors which are of light weight, small size, high precision and have a wide frequency response, a chance is offered to the discontiguous machinery failure diagnosis[6]. From the machine failure diagnosis study point of view, the discontiguous machinery fault detection technology may not be new. For example, the machinery noise analysis and acoustic emission detection technology are also discontiguous fault detection technology. There are some successful examples of machinery fault diagnosis by using machinery noise analysis and acoustic emission technology, but these methods are easily influenced by ambient noise under actual factory running conditions[11].

In this paper, we are going to find a new approach for discontiguous machinery fault diagnosis techniques. The failure detection ability is discussed for using laser displacement sensors under the two kinds of sensor transfer time, different failure sizes and different shaft speeds. Experiment results show that it has the ability to solve the problem of no-contact detecting machine failure.

EXPERIMENTAL SET-UP AND CONDITIONS: The experimental set-up is shown in Fig. 1. The experimental set-up consisted of a high speed motor mounted on a 20 cm-thick

base plate. The motor was connected to the driven shaft on which the auxiliary bearings and test bearings were housed by a belt and pulley system. The motor was controlled by an alternating governor system. The test load was provided by a hydraulic cylinder.

The vibration displacement was measured by a special laser displacement sensor from the outer surface of bearing outer race or the top of bearing housing. The feature of laser sensor will be discussed later. The signal passed through an A/D board to the computer, where signals were processed.

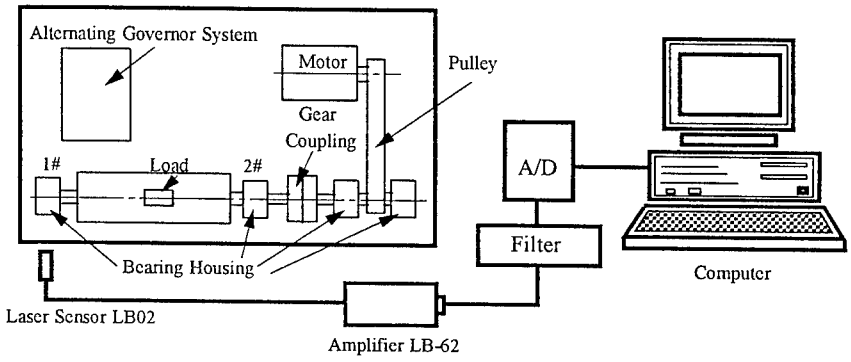


Fig. 1 Experimental Set-up and Data Acquisition

The Feature of Laser Displacement Sensors: 1) Features of Frequency Response: The laser sensor type is LB02, amplifier type is LB-62. Its linear frequency range of response is 0-3 kHz, its maximum frequency range of response can be over 10 kHz. The frequency range of response is relative to the transfer speed of the laser sensor. It was divided into three areas according to the three transfer speeds of the laser sensor. The relationship of the laser sensor's resolution and the transfer speed is shown in Fig. 2. **2) Resolution of Displacement Sensor:** The resolution of the laser sensor is 2-50 μm for vibration displacement. The resolution is relative to the transfer speed of the laser sensor. It was also divided into three areas according to the three transfer speeds of the laser sensor, they are 2 μm to 60 ms, 15 μm to 2 ms and 50 μm to 0.15 ms. Fig. 2 also shows that it has a low resolution when the transfer speed is high, there is an inverse ratio relation between the resolution and the transfer speed. Attention should be paid to this important feature when applying the laser sensor to measurement vibration displacement. **3) Influence of the Angled Object:** Since there is an angle on the bearing outer race or bearing housing surface, so it is an important factor for detecting the failure of bearings by using a laser sensor which can detect angle range. It has only a small influence on the angled object when the reflection angle of the object is in the range $\pm 30^{\circ}$. Thus, the vibration displacement of bearings can be measured by using a laser displacement sensor on the bearing housing or outer race. Fig. 3 shows the result.

Placement of the Measurement Point: The vibration displacement is different from the vibration acceleration, its losses are greater when the signals pass once through the contact interface. According to the feature of the laser displacement sensor, there are two measurement placements to be selected in the experiment. One point is from the outer race of bearing, the other one is from the bearing housing (both in 1# point of Fig. 1).

Experimental Conditions: The experimental self-aligning bearings are a bearing type of NSK 22207. These bearings could be artificially localized defects induced by an electric-discharge machine. The bearing failure occurs with defects in the outer race and roller. The experimental data includes one failure size type of self-aligning bearing under two kinds of shaft speed and one load. The test conditions are shown in Table 1.

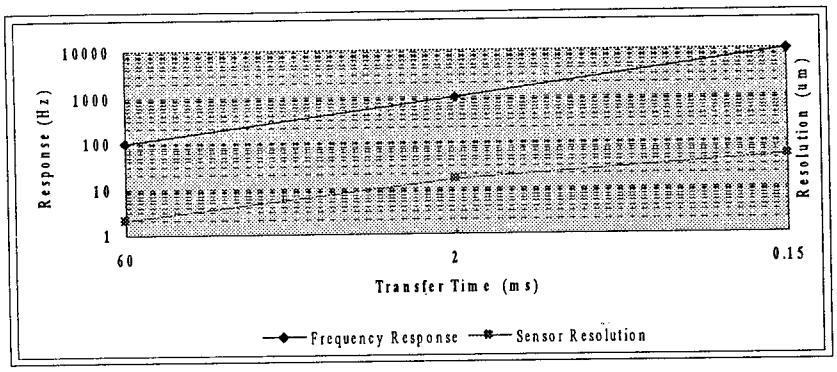


Fig. 2 The Relationship of Transfer Time, Frequency Response and Resolution

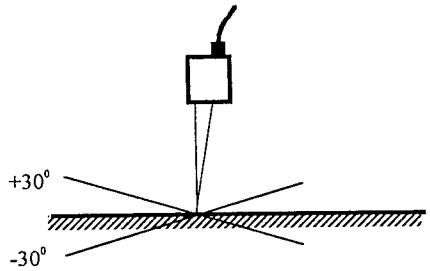


Fig. 3 Measurement Range of Angled Object

Table 1 The Experimental Conditions

Conditions	Values	Outer Race Frequency
Defect Type 1	8.3x1x1.5 (mm)	——
Defect Type 2	3x1x0.5 (mm)	——
Shaft Speed 1	1000 rpm	27 Hz
Shaft Speed 2	500 rpm	53 Hz
Shaft Speed 3	250 rpm	106 Hz
Loads	5 kN	——

The sampling frequency is 2 and 5 kHz. A time domain data set contains 3 samples, a sample has 1024 points. Spectra are averaged over 3 samples with a HANNING window. A figure of spectrum was drawn by using 512 points. The different resolutions (15 and 50 μ m) and different transfer times (0.15 and 2 ms) of laser sensor were selected in this study.

DISCUSSION OF RESULTS: It contains two sampling frequencies when using the A/D board as a data transfer device, and it uses different resolutions of laser sensor to detect bearing vibration. If the vibration displacement variation period produced by localized defects can be detected by the different resolutions and different transfer speeds of laser sensors, then results of bearing running conditions can be obtained by signal process technology.

Vibration Displacement Signals from the Surface of Outer Race:

Fig. 4 to Fig. 9 show the spectrum analysis results of vibration signals from the surface of outer race under the two failure sizes and three shaft speeds. The resolution of laser sensor is 15 μ m (transfer time of laser sensor is 2 ms).

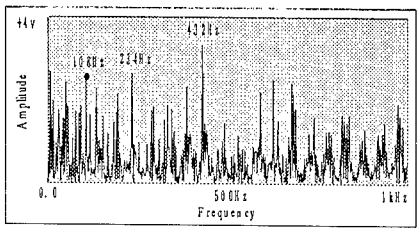


Fig. 4 Outer Race Defect 1 Under Shaft Speed 1

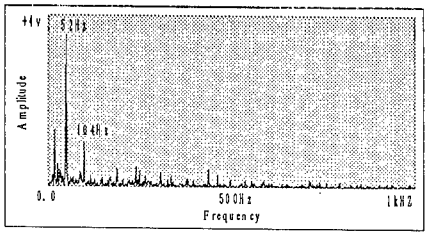


Fig. 5 Outer Race Defect 1 Under Shaft Speed 2

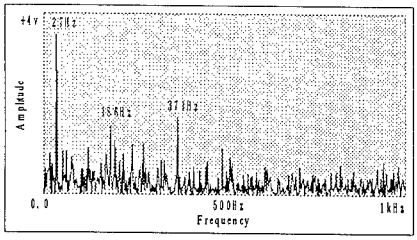


Fig. 6 Outer Race Defect 1 Under Shaft Speed 3

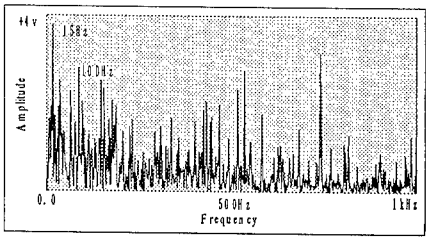


Fig. 7 Outer Race Defect 2 Under Shaft Speed 1

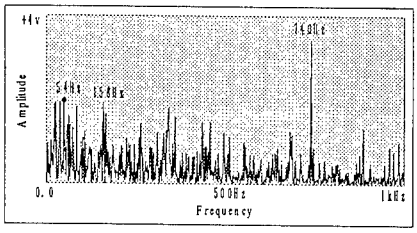


Fig. 8 Outer Race Defect 2 Under Shaft Speed 2

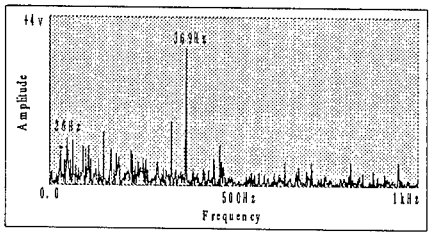


Fig. 9 Outer Race Defect 2 Under Shaft Speed 3

Figs. 4, 5 and 6 show outer race fault frequencies of 108 Hz, 52 Hz and 27 Hz, respectively. In Figs. 5 and 6 failure frequency of the outer race is more clear than in Fig. 4. The result shows that there is a great amount of frequency peaks in the vibration displacement signal when the bearing is running at high speed, so the failure of bearing is easily identified when the bearing is running at low speed.

From Fig. 7, Fig. 8 and Fig. 9, we can obtain a similar spectrum analysis result when the failure size is small, but the peaks of failure frequency are not clear in these figures under the same shaft speed and sensor resolution. The result shows that failure size influences the ability to detect failure using a laser displacement sensor.

Now we investigate the features of time domain statistical values under different shaft speeds, failure sizes, and the same sensor resolution conditions. Figs. 10 and 11 show 5 types of the characteristic statistical parameter values. They are shape factor, crest factor, impulse factor, clearance factor and kurtosis value. Comparing Figs. 10 and 11 show that the time domain statistical values to some extent increase when defect size is bigger under the same shaft speed and sensor resolution conditions. We also know from Figs. 10 and 11 that the shaft speeds have little influence over the time domain statistical value.

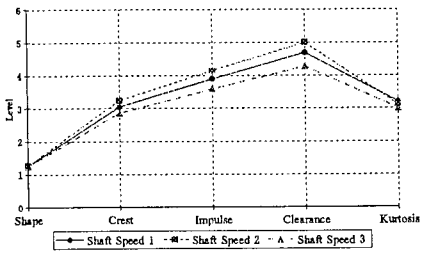


Fig. 10 Time Domain Statistical Values of Defect 1 Under Resolution 15 μ m rpm

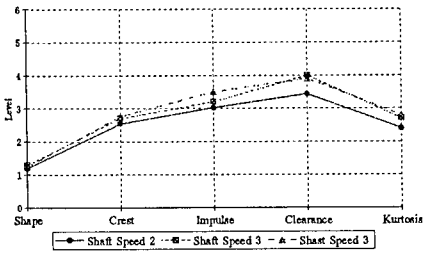


Fig. 11 Time Domain Statistical Values of Defect 2 Under Resolution 15 μ m rpm

Vibration Displacement Signals from Bearing Housing: Since vibration displacement is from the bearing housing, there are greater losses in the measurement signal when the vibration displacement signals pass once through the contact interface. This is important research work that failure of bearing was detected by vibration displacement signal from bearing housing. In order to compare the result with signals from the outer race, the same processing method was applied in this paragraph.

1) The Outer Race Defect 1:

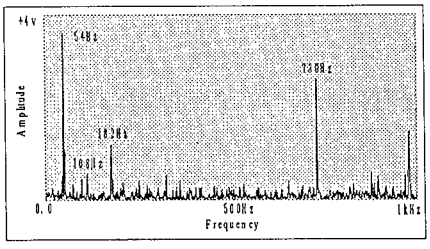


Fig. 12 Sensor Resolution 50 μ m Under Shaft Speed 1

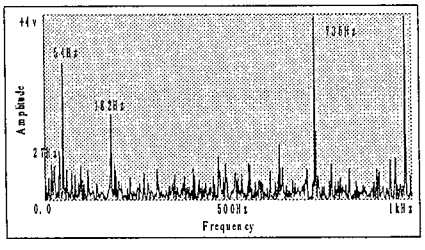


Fig. 13 Sensor Resolution 50 μ m Under Shaft Speed 2

Making a simple comparison between these spectrum, the peaks of failure feature frequency we can find in Fig.12, Fig.13 and Fig.14 under sensor resolution is 15 μ m. Sensor resolution 15 μ m shows better results in Fig. 15, Fig. 16 and Fig. 17. Comparing sensor resolution 15 μ m and 50 μ m, the high sensor resolution case has a better ability of failure detection than the low sensor resolution. We have to point out that 54 Hz frequency is very obvious in the Figs. 12 and 13, and also obvious in Figs. 15, 16 and 17, so it will be difficult to identify the bearings failure type. Although the peaks of failure feature frequency are not easy to identify in these spectrum, they show that bearings failure can be detected by using the vibration displacement from the bearing housing under the different sensor resolutions and different shaft speeds.

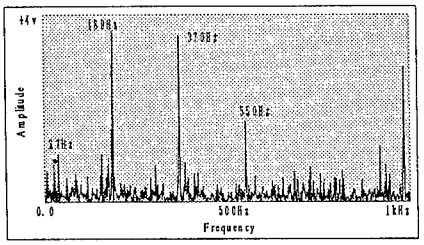


Fig. 14 Sensor Resolution 50 μ m Under Shaft Speed 3

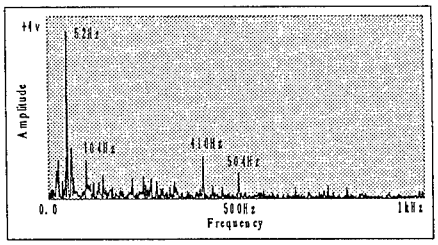


Fig. 15 Sensor Resolution 15 μ m Under Shaft Speed 1

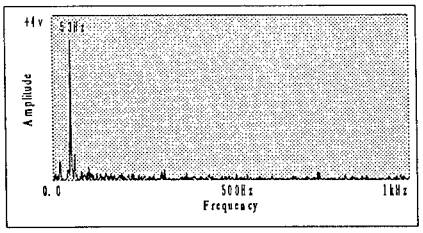


Fig. 16 Sensor Resolution 15 μ m Under Shaft Speed 2

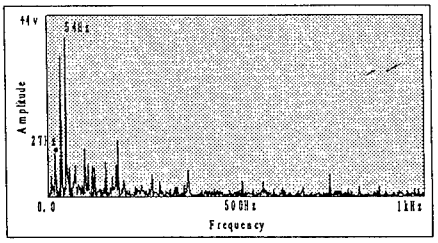


Fig. 17 Sensor Resolution 15 μ m Under Shaft Speed 3

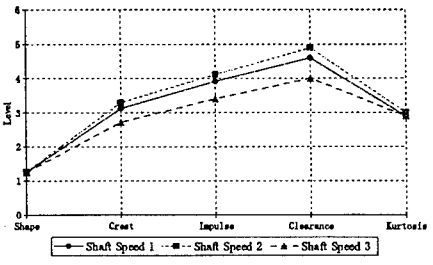


Fig. 18 Time Statistical Values of Sensor Resolution 50 μ m

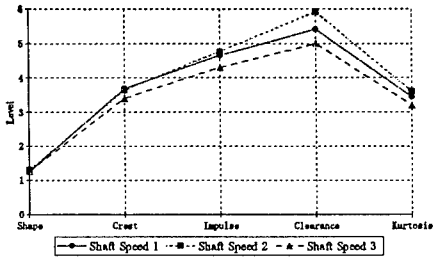


Fig. 19 Time Statistical Values of Sensor Resolution 15 μ m

Let's now look at feature of the time domain statistical value. Figs. 18 and 19 show a similar result to Figs. 10 and 11, where the shaft speeds have only a little influence over time domain statistical values. Comparing Figs. 18 and 19, we see to some extent an increase of statistical values in Fig. 19 when the sensor resolution is 15 μ m, it also shows that the sensor resolution is very important for failure detection under the same test condition.

2) The Outer Race Defect 2: The failure frequency of the outer race can be found in Fig. 20, Fig. 21 and Fig. 22, where the failure size is small. Since there are more peaks of frequency in these spectrum, the peaks of failure feature frequency become hard to find. The result shows that the failure size will effect the ability of failure identification, although the sensor resolution is the same. The time domain statistical values show to some extent a decrease when the failure size is small, the result can be obtained by comparing Figs. 11 and 19.

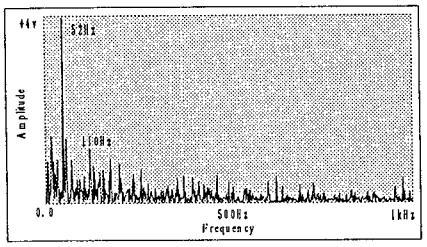


Fig. 20 Sensor Resolution 15 μ m Under Shaft Speed 1

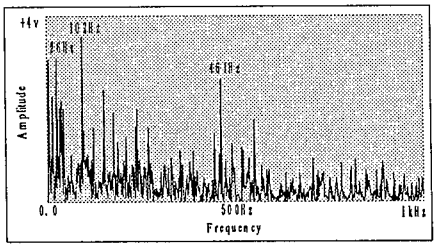


Fig. 21 Sensor Resolution 15 μ m Under Shaft Speed 2

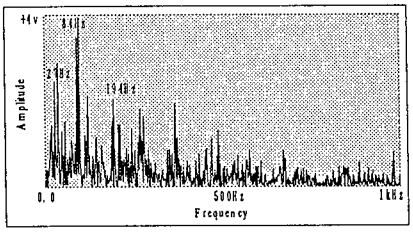


Fig. 22 Sensor Resolution 15 μ m Under Shaft Speed 3

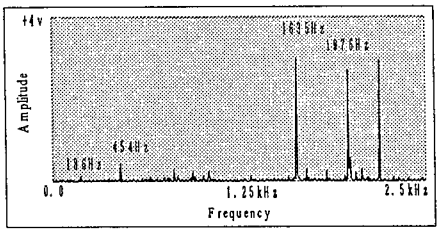


Fig. 23 Sensor Resolution 15 μ m Under Shaft Speed 1 (Sampling 5 kHz)

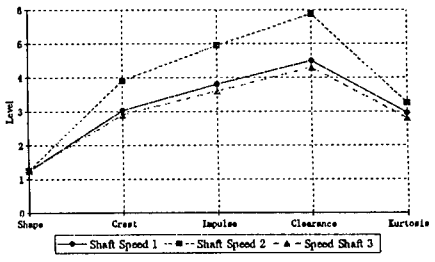


Fig. 24 Time Statistical Values of Sensor Resolution 15 μ m

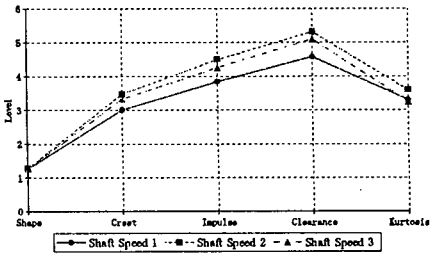


Fig. 25 Time Statistical Values of Sensor Resolution 15 μ m (Sampling 5 kHz)

3) The Sampling Frequency 5 kHz: It is difficult to find obvious peaks of failure frequency in Fig. 23, but through making a comparison of Figs. 24 and 25, we can see the time domain statistical values have little change. The result shows that the sampling frequency will influence the ability of failure detection using a laser displacement sensor, but it still can get similar results when using the time statistical values as failure detection means.

The above results show that the spectrum analysis and time domain statistical values have the ability to detect bearing failures whether the displacement signals of vibration are from the outer race of bearing or from bearing housing. It shows a good ability of failure detection under the shaft speed 2 and the sensor resolution 15 μ m.

CONCLUSION: It has been proved that the displacement analysis method is an effective method to identify faults whether the displacement comes from the surface of outer race or from the top of the bearing housing, whether at a high shaft speed or at a low shaft speed. Certainly, it has more actual value of application to detect the signal from the bearing housing.

The results of the spectrum analysis will be influenced by different sampling frequencies, but the time domain statistical values are stable for two sampling frequencies. The failure sizes will effect the ability of failure detection, but the shaft speed will have little influence on the ability of failure detection.

These results have shown that the laser displacement sensor has the ability to solve the problem of discontiguous failure detection. A further study is necessary for the monitoring and diagnosis of bearing failure using laser sensor under different bearing work conditions.

REFERENCES

[1] Yimin Shao, Kikuo Nezu, "Feature Extraction of Machinery Diagnosis Using Neural Network", Proceedings of IEEE International Conference on Neural Networks, Vol. 1, 459-464, Perth, Australia, November, 1995.

[2] Yimin Shao, Kikuo Nezu, "An On-Line Monitoring and Diagnostic Method of Rolling Element Bearing With AI", Trans. of the Society of Instrument and Control Engineers, Vol. 32, No. 8, 1287/1293, 1996.

[3] Yimin Shao, Kikuo Nezu, " Self-aligning Roller Bearing Fault Detection Using Asynchronous Adaptive Noise Cancelling ", Proceedings of JOAP & MFPT, No. 50, p633-640, Mobile, Alabama, U.S.A., April 22-26, 1996.

[4] Karl Barthel, "The Shock Pulse Method for Measuring the Condition of Antifriction Bearing", Tappi, Vol. 60, No. 8, Aug. 1977.

[5] G. K. Chaturvedi, D. W. Thomas, "Bearing fault detection using adaptive noise cancelling", Trans. of the ASME Journal of Mechanical Design, Vol. 104, 280- 289, 1982.

[6] Yimin Shao, Kikuo Nezu, " Bearing Fault Detection Using Laser Displacement Sensor " Proceedings of SICE'96, 1069-1072, Tottori, July 1996, Japan.

[7] D. Dyer, R. M. Stewart, " Detection of Rolling Element Bearing Damage by Statistical Vibration Analysis", Trans. of the ASME Journal of Mechanical Design, Vol. 100, 229-235, April, 1978.

[8] J. I. Taylor, "Identification of Bearing Defects by Spectra Analysis", Trans. of the ASME Journal of Mechanical Design, Vol. 102, 198-204, April, 1980.
[9] J. Xu, H. Peeken, " Failure Diagnosis System Using Fuzzy Logic", The 1989 ASME Design Technical Conference-12th Biennial Conference on Mechanical Vibration and Noise, De-Vol. 18-3, 18-4, 18-5, 93-99, Montreal Quebec Canada, 17-21, 1989.
[10] Simon G. Braun, "The Signature Analysis of Sonic Bearing Vibration", IEEE Trans. on Sonic and Ultrasonic, Vol. Su-27, No. 6, November 1980.
[11] Kexing Cheng, etc., " Machinery Failure Detection and Diagnosis Technique ", Beijing , China, 1991.
[12] I. E. Alguindigue, etc., "Monitoring and Diagnosis of Rolling Element Bearings Using artificial Neural net works", IEEE Trans. on Industrial Electronics, Vol. 40, No. 2 April, 1993.

RSAFP

NUMERICAL AND ANALYTICAL METHODS

Chair: **Omar Es-Said**
Loyola Marymount University

THE DYNAMICS STRESS INTENSITY FACTOR OF SURFACE AND INTERNAL CRACK SUBJECTED TO OBLIQUELY SHEAR WAVES

I-Chung Weng and Zhiming Ji
Department of Mechanical Engineering
New Jersey Institute of Technology
Newark, New Jersey 07102, USA.

Abstract: By using the dislocation model for the surface crack and images for the free surface, an analytical solution for interaction between a screw dislocation near a mode III surface crack subjected to the dynamic antiplane stress has been derived. The change of the crack extension force as the surface crack propagation towards the dislocation is presented in this paper. Without the applied stress, the dislocation stress field tends to drag the surface crack to the dislocation where the strain energy can be relaxed. Under the periodic dynamic stress, the surface crack can be repelled by the dislocation depending upon the direction of the applied stress and wave number and the Burgers vector of the dislocation. The effects of the wave number, the input incident angle and the dislocation on the stress intensity factor are also presented. The results showed this problem can be reduced to a crack embedded in an infinite solid when the distance between the dislocation and the free surface is ≥ 20. This method is applied to find out the dynamic stress intensity factor of an internal crack near a free surface under SH loadings.

Key Words: Dislocation model; screw dislocation; crack propagation; Burgers vector; wave number; input incident angle; stress intensity factor

INTRODUCTION: Dislocation around a crack tip plays an important role in the ductility of the material. It has been observed in silicon wafer that the dislocations near the free surface can be generated by introducing surface damage followed by a proper heat treatment. The surface damage, caused by grinding, scratching, etc., on brittle material usually introduces surface microcrack along with dislocations. It is believed that the dislocations are generated by surface microcracks during the heat treatment. It is therefore important to understand the interaction between a surface crack and a dislocation. Louat [1] was the first to study the interaction between the dislocation and the finite crack. Smith [2] showed that an edge dislocation in front of a crack tip can enhance the growth of cracks of Griffith sizes, and he also formulated the dislocation distribution for a finite crack that interacts with coplanar screw dislocations. For a semi-infinite crack in the infinite medium, the image force components on the dislocation along the slip [3] and climb directions [4] have been solved by the thermodynamic method. Chu [5] solved the coplanar screw dislocation and sharp surface crack interaction. Lee [6] also studied the same problem and compared the dislocation distributions in the crack, the total Burger vector, and the stress intensity factor in the models of (Li , 7) and (Louat , 1). Li [7] used dislocation modeling of the crack tip stress field to study the nucleation of dislocation near the tip of the crack. The general problem

of the interaction of a dislocation with a surface notch has been studied by Warren [8]. However, due to the complexity of the potential he obtained, only an approximate solution for the force on the dislocation was reached. The dynamic interaction between the general parallel screw dislocations and the surface crack still has not been solved. This work considers the case of the interaction of a screw dislocation with a mode III surface crack subjected to harmonic periodic stress. By using the conformal mapping technique, the stress along the crack is expressed as a system of singular integral equations, which contain Bessel functions and the distribution densities, the Burgers vector, and the phase lags of the dislocations. From the Simpson integration method, we expressed the unknowns in the forms of the Chebyshev Polynomials and solved them with Galerkin method.

PROBLEM DEFINITION: The problem considered in this study is as follows: A surface crack of unit length lying in the xz plane with an infinite z dimension is shown in Fig. 1. A screw dislocation parallel to the z axis is situated at $(x_0, 0)$ in the xz plane (Fig. 2). Both the crack and the dislocation are subjected to horizontal polarized shear waves (SH waves). The screw dislocation has a Burgers vector b. The problem is to calculate the stress intensity factor at the crack tip due to both the applied shear waves and the screw dislocation. We represent the surface crack with an array of continuous distribution of screw dislocations [6], all parallel to the z axis. Owing to the nature of geometry, the stress on the free surface is zero. The problem is now changed to simply a finite crack of length 2 interacting with two screw dislocations in an infinite medium as shown in Fig. 2. For the case of harmonic shear waves (SV, SH) impinging on a semi-infinite crack lying along x-axis, the anti-plane displacement component of the incident field at point (x, y) can be expressed as [9]:

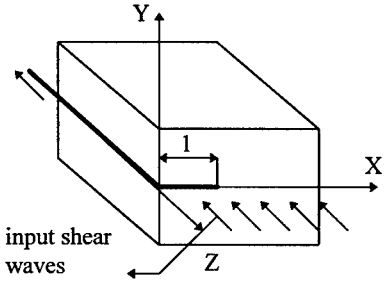


Figure 1 The surface crack under the SH waves

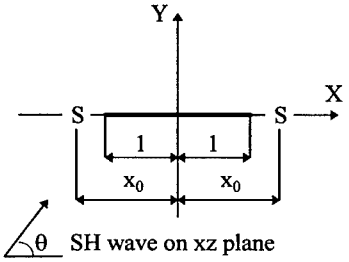


Figure 2 A positive screw dislocation, its image and a finite crack of length 2.

$$w_k = w_0\exp\{-i[\alpha(x\cos\theta+y\sin\theta)-\omega t]\}$$
$$= w_0\cos[\alpha(x\cos\theta+y\sin\theta)-\omega t]$$
$$-iw_0\sin[\alpha(x\cos\theta+y\sin\theta)-\omega t] \quad (1)$$

where w_0 and ω are the amplitude and frequency of the applied wave, w_k is the anti-plane displacement component, θ is the angle of incidence measured from the x-axis, and α is the wave number. Let μ and ρ be respectively the shear modulus and the mass density, then the shear wave velocity is the $(\mu/\rho)^{1/2}$ and the wave number $\alpha = \omega/c$. The displacement w_k can be decomposed into real (SV) and imaginary (SH) components

which are parallel to x-axis and y-axis respectively. This study concentrates on the stress intensity factor of mode III. That is, only the SH wave problem is considered. The anti-plane stress under SH waves can be derived if we denote the imaginary component of Eq. (1) as w^*, then

$$\sigma^* = -\mu\partial w^*/\partial y = \sigma_0 \sin\theta\cos[\alpha(x\cos\theta+y\sin\theta)-\omega t] \quad (2)$$

where $\sigma_0 = w_0\mu\alpha$. When the input shear waves meet the dislocation, these waves are reflected due to the inhomogeneity of dislocation density. Previous works on this subject showed that if the input stress waves are SH waves, the reflected waves would also be SH waves. This result is equivalent to that produced by reflection from a crack. Thus, the deformation at a crack can be modeled as a continuous distribution of screw dislocation under SH waves. From Eshelby's study [10], the total displacement at point (x, y) when a shear wave is incident normally on the screw dislocation is expressed as:

$$d(r) = \frac{\alpha b u}{4}\left(\frac{2}{\pi\alpha r}\right)^{1/2}\cos[\omega t+p(r)]\sin\theta \quad (3)$$

where $r = (x^2+y^2)^{1/2}$, p(r) is the phase lag, u is the amplitude of the screw dislocation and b is the Burger's vector. Eshelby also showed that the scattered wave may be represented by Bessel functions of order 1. The anti-plane displacement is:

$$w(r) = A\sin\theta\{J_1(\alpha|r|)\cos[\omega t+p(r)] + Y_1(\alpha|r|)\sin[\omega t+p(r)]\} \quad (4)$$

Another approach related to the interaction of screw dislocations and sound waves was studied by Nabarro [11]. He derived the total displacement as follows:

$$d(r) \approx \left(\frac{2}{\pi\alpha r}\right)^{1/2} A\cos[\omega t+p(r)]\sin\theta \quad (5)$$

Comparing the Eqs. (3) and (5), we get $A \approx \alpha b/4$. Inserting Eq. (4) into $\tau = -\mu\partial w(x)/\partial y$, the anti-plane stress wave released along y = 0 plane is :

$$\sigma_d(x) = B\alpha^2\{[J_0(\alpha|x|) + J_2(\alpha|x|)]\cos[\omega t+p(x)] + [Y_0(\alpha|x| + Y_2(\alpha|x|)]\sin[\omega t+p(x)]\} \quad (6)$$

where $B = -bu\mu/8$, and J_0 and J_2 are the zeroth order and the second order of the first kind of Bessel function respectively. Similarly, Y_0 and Y_2 are the zeroth order and the second order of the second kind of Bessel function. From Eq. (2), stress σ^* along the crack surface (y = 0 plane) is $\sigma^* = \sigma_0\sin\theta\cos(\alpha x\cos\theta-\omega t)$. Since the total surface traction should be zero along the crack surface, we have

$$\int_{-1}^{1} D(s)\, B\alpha^2\{[J_0(\alpha|x-s|) + J_2(\alpha|x-s|)]\cos[\omega t+p_1(s)]+[Y_0(\alpha|x-s|) +Y_2(\alpha|xs|)]\sin[\omega t+p_1(s)]\}ds$$
$$+ B\alpha^2\{[J_0(\alpha|x-x_0|) + J_2(\alpha|x-x_0|)]\cos[\omega t+p_2(x_0)]+ [Y_0(\alpha|x-x_0|) + Y_2(\alpha|x-x_0|)]\sin[\omega t+p_2(x_0)]$$
$$- [J_0(\alpha|x+x_0|) + J_2(\alpha|x+x_0|)]\cos[\omega t+p_2(x_0)]- [Y_0(\alpha|x+x_0|) + Y_2(\alpha|x+x_0|)]\sin[\omega t+p_2(x_0)]\}$$
$$= -\sigma_0\sin\theta\cos(\alpha x\cos\theta-\omega t) \quad (7)$$

where the $p_1(s)$ and $p_2(x_0)$ are phase lags of the surface crack modeled by a continuous distribution of screw dislocation and the dislocation at $(x_0, 0)$ respectively. Expanding Eq. (7), we get the following two equations:

$$\int_{-1}^{1} D(s)\,B\alpha^2[Y\cos p_1(s) - J\sin p_1(s)]ds$$
$$= -\sigma_0 \sin\theta \sin(\alpha x\cos\theta) + B\alpha^2(\eta-\varepsilon)\cos p_2(x_0) + B\alpha^2(\chi-\xi)\sin p_2(x_0) \quad (8a)$$
$$\int_{-1}^{1} D(s)\,B\alpha^2[J\cos p_1(s) + Y\sin p_1(s)]ds$$
$$= -\sigma_0 \sin\theta \cos(\alpha x\cos\theta) + B\alpha^2(\eta-\varepsilon)\sin p_2(x_0) + B\alpha^2(\xi-\chi)\cos p_2(x_0) \quad (8b)$$

where $J = J_0(\alpha|x-s|) + J_2(\alpha|x-s|)$, $Y = Y_0(\alpha|x-s|) + Y_2(\alpha|x-s|)$, $\chi = J_0(\alpha|x-x_0|) + J_2(\alpha|x-x_0|)$, $\varepsilon = Y_0(\alpha|x-x_0|) + Y_2(\alpha|x-x_0|)$, $\xi = J_0(\alpha|x+x_0|) + J_2(\alpha|x+x_0|)$, $\eta = Y_0(\alpha|x+x_0|) + Y_2(\alpha|x+x_0|)$. On the right side of Eqs. (8a, 8b) includes three parts: where the first part is due to the applied SH waves. The second and third are due to the screw dislocation at $(x_0, 0)$ and its image respectively. There are two unknown functions $D(s)$ and $p_1(s)$ on the left side of Eqs. (8a, 8b). We replace them with another two functions, $A_1(s) = D(s)B\cos p_1(s)$ and $A_2(s) = D(s)B\sin p_1(s)$. Since there is singularity in Y when the value of x approaches to s, we can separate Y into singular and regular parts. Eqs. (8a, 8b) are expressed as follows:

$$-4/\pi \int_{-1}^{1} A_1(s)/(x-s)^2 ds + 2\alpha^2/\pi \int_{-1}^{1} A_1(s) J\ln(\alpha|x-s|)ds + \int_{-1}^{1} \alpha^2 [A_1(s)f(x,s) - A_2(s)J]ds$$
$$= -\sigma_0 \sin\theta \sin(\alpha x\cos\theta) + B\alpha^2(\eta-\varepsilon)\cos p_2(x_0) + B\alpha^2(\chi-\xi)\sin p_2(x_0) \quad (9a)$$
$$-4/\pi \int_{-1}^{1} A_2(s)/(x-s)^2 ds + 2\alpha^2/\pi \int_{-1}^{1} A_2(s) J\ln(\alpha|x-s|)ds + \int_{-1}^{1} \alpha^2 [A_2(s)f(x,s) + A_1(s)J]ds$$
$$= -\sigma_0 \sin\theta \cos(\alpha x\cos\theta) + B\alpha^2(\eta-\varepsilon)\sin p_2(x_0) + B\alpha^2(\xi-\chi)\cos p_2(x_0) \quad (9b)$$

where $f(x,s) = -1/\pi + 2/\pi[(\gamma - \ln 2)J_0(\alpha|x-s|) - \ln 2 J_2(\alpha|x-s|)] - 2/\pi \sum_{j=0}^{\infty} (-1)^j (\alpha/2)^{2j+2}(x-s)^{2j+2}$ $\left[\frac{\psi(j+1)+\psi(j+3)}{2\,j!(2+j)!} - \frac{\psi(j+2)+\gamma}{[(j+1)!]^2}\right]$, with the Euler constant $\gamma \cong .577215665$ and $\psi(j) = -\gamma + \sum_{k=1}^{j-1}(1/k)$, $j > 1$. The dynamic stresses around a small circle centered at the crack tip ($r_2 \to 2a$ and $\phi_2 \to 0$) under shear waves can be expressed as [12]:

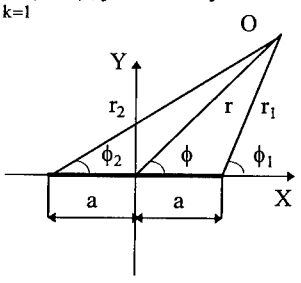


Figure 3 Definitions of r_i and ϕ_i

$$\tau_{yz} = \frac{K_{III}}{\sqrt{2r_1}}\cos(\phi_1/2),\ \tau_{xz} = -\frac{K_{III}}{\sqrt{2r_1}}\sin(\phi_1/2) \quad (10a, b)$$

where r_i and ϕ_i ($i = 1, 2$) are shown in Fig. 3. K_{III} is the stress intensity factor of tearing mode. That means the stress intensity factors at the tip of the crack rely on the existence of the $r^{1/2}$ stress singularity. We expect that, according to the Simpson integration method, $A_1(s)$ and $A_2(s)$ contain the term $(1-s^2)^{1/2}$. Since Eqs. (9a, 9b) contain orthogonal polynomials, we choose to express $A_1(s)$ and $A_2(s)$ in the following forms:

$$A_1(s) = (1-s^2)^{1/2} \sum_{n=0}^{\infty} a_n U_n(s),\ A_2(s) = (1-s^2)^{1/2} \sum_{n=0}^{\infty} b_n U_n(s) \qquad (11a, b)$$

where $U_n(s)$ is the second kind of Chebyshev Polynomial [13]. Introducing the Galerkin method into Eqs. (9a, 9b) and substituting $A_1(s)$ and $A_2(s)$ from Eqs. (11a, 11b) into Eqs. (9a, 9b), we have:

$$\sum_{n=0}^{\infty} a_n \{-4/\pi \int_{-1}^{1} U_m(x)(1-x^2)^{1/2} [\int_{-1}^{1} \frac{U_n(s)}{(x-s)^2}(1-s^2)^{1/2} dsdx + \alpha^2 [\int_{-1}^{1} \frac{2}{\pi} J \ln(\alpha|x-s|) +$$

$$f(x,s)]U_n(s)(1-s^2)^{1/2} dsdx] - \sum_{n=0}^{\infty} b_n \alpha^2 \int_{-1}^{1} U_m(x)(1-x^2)^{1/2} \int_{-1}^{1} J U_n(s)(1-s^2)^{1/2} dsdx$$

$$= -\int_{-1}^{1} U_m(x)(1-x^2)^{1/2} [\sigma_0 \sin\theta \sin(\alpha x \cos\theta)]dx + \int_{-1}^{1} U_m(1-x^2)^{1/2} B\alpha^2 (\eta-\varepsilon)\cos p_2(x_0)dx$$

$$+ \int_{-1}^{1} U_m(1-x^2)^{1/2} B\alpha^2 (\chi-\xi)\sin p_2(x_0)dx \qquad (12a)$$

$$\sum_{n=0}^{\infty} b_n \{-4/\pi \int_{-1}^{1} U_m(x)(1-x^2)^{1/2} [\int_{-1}^{1} \frac{U_n(s)}{(x-s)^2}(1-s^2)^{1/2} dsdx + \alpha^2 [\int_{-1}^{1} \frac{2}{\pi} J \ln(\alpha|x-s|)$$

$$+ f(x,s)]U_n(s)(1-s^2)^{1/2} dsdx\} + \sum_{n=0}^{\infty} a_n \alpha^2 \int_{-1}^{1} U_m(x)(1-x^2)^{1/2} \int_{-1}^{1} J U_n(s)(1-s^2)^{1/2} dsdx$$

$$= -\int_{-1}^{1} U_m(x)(1-x^2)^{1/2} [\sigma_0 \sin\theta \cos(\alpha x \cos\theta)]dx + \int_{-1}^{1} U_m(1-x^2)^{1/2} B\alpha^2 (\eta-\varepsilon)\sin p_2(x_0)dx$$

$$+ \int_{-1}^{1} U_m(1-x^2)^{1/2} B\alpha^2 (\xi-\chi)\cos p_2(x_0)dx \qquad (12b)$$

where m = 0, 1, 2, 3, ..., M, ...∞ and n = 0, 1, 2 3, ..., N, ..., ∞. The integrals in Eqs. (12a, 12b) are evaluated with the help of Chebyshev Polynomials. Since U_n does not vanish when n approaches ∞, a_n and b_n should converge to zero as n approaches ∞. Therefore we assume that the problem can be approximated with the following series:

$$\sum_{n=0}^{N} [(A_{mn} + B_{mn})a_n - C_{mn}b_n] = \eta_m + f_m \cos p_2(x_0) + g_m \sin p_2(x_0) \quad (m = 0, 1, ..M) \qquad (13a)$$

$$\sum_{n=0}^{N} [C_{mn}a_n + (A_{mn} + B_{mn})b_n] = \nu_m + f_m \sin p_2(x_0) - g_m \cos p_2(x_0) \quad (m = 0, 1, ..M) \qquad (13b)$$

$$A_{mn} = -4/\pi \int_{-1}^{1} U_m(x)(1-x^2)^{1/2} \int_{-1}^{1} \frac{U_n(s)}{(x-s)^2}(1-s^2)^{1/2} dsdx = \begin{cases} 0 & m \neq n \\ 2(n+1)\pi & m = n \end{cases}$$

$$B_{mn} = \alpha^2 \int_{-1}^{1} U_m(x)(1-s^2)^{1/2} \int_{-1}^{1} [\frac{2}{\pi} J \ln(\alpha|x-s|) + f(x,s)]U_n(s)(1-s^2)^{1/2} dsdx$$

$$C_{mn} = \alpha^2 \int_{-1}^{1} U_m(x)(1-x^2)^{1/2} \int_{-1}^{1} J U_n(x)(1-s^2)^{1/2} dsdx,\ g_m = \int_{-1}^{1} U_m(1-x^2)^{1/2} B\alpha^2 (\chi-\xi)dx$$

$$\eta_m = -\int_{-1}^{1} U_m(x)(1-x^2)^{1/2} [\sigma_0 \sin\theta \sin(\alpha x \cos\theta)]dx,\ f_m = \int_{-1}^{1} U_m(1-x^2)^{1/2} B\alpha^2 (\eta-\varepsilon)dx$$

$$\nu_m = -\int_{-1}^{1} U_m(x)(1-x^2)^{1/2}[\sigma_0 \sin\theta\cos(\alpha x\cos\theta)]dx .$$

Eqs. (13a, 13b) contain 2(N+1) unknown coefficients a_n and b_n with 2(M+1) equations. The above equations can be put into matrix form as:

$$(A+B)\vec{a} - C\vec{b} = \vec{\eta} + \vec{f}\cos p_2(x_0) + \vec{g}\sin p_2(x_0) \quad (14a)$$

$$C\vec{a} + (A+B)\vec{b} = \vec{\nu} + \vec{f}\sin p_2(x_0) - \vec{g}\cos p_2(x_0) \quad (14b)$$

where $\vec{a} = [\, a_0 , a_1 , a_2 ,...,a_N \,]^T$ and $\vec{b} = [\, b_0 , b_1 , b_2 ,...,b_N \,]^T$. Matrices A, B, C, and vectors $\vec{\eta}, \vec{\nu}$ and $\vec{f}, \vec{g}$ can be found once M and N are selected. Since all quantities in Eqs. (13a, 13b) are constants except a_n and b_n, we deal with them as linear equations. Separating them by superposition method, Eqs. (14a, 14b) become:

$$(A+B)\vec{a}' - C\vec{b}' = \vec{\eta}, \; C\vec{a}' + (A+B)\vec{b}' = \vec{\nu} \quad (15a, b)$$

$$(A+B)\vec{a}'' - C\vec{b}'' = \vec{f}\cos p_2(x_0) + \vec{g}\sin p_2(x_0) \quad (16a)$$

$$C\vec{a}'' + (A+B)\vec{b}'' = \vec{f}\sin p_2(x_0) - \vec{g}\cos p_2(x_0) \quad (16b)$$

where $\vec{a} = \vec{a}' + \vec{a}''$ and $\vec{b} = \vec{b}' + \vec{b}''$. To solve for a_n and b_n (n = 0, 1, ..., N), let M = N. Our problem is now to find positive integers M and N such that a_n and b_n converge to zero. As a first approximation, let M and N equal to 9. Evaluating a'_n and b'_n in Eqs. (15a, 15b), we find out the values of a'_n and b'_n would converge to zero if M, N > 5. The results are listed in Table I for $\alpha = 1$ and $\theta = \pi/2$. Let

$$\vec{f}\cos p_2(x_0) + \vec{g}\sin p_2(x_0) = \vec{R}\cos\varphi, \; \vec{f}\sin p_2(x_0) - \vec{g}\cos p_2(x_0) = \vec{R}\sin\varphi \quad (17a, b)$$

where $0 \le \varphi \le 2\pi$. We now have

$$\vec{R} = (\vec{f}^2 + \vec{g}^2)^{1/2} \quad (18)$$

The corresponding a''_n and b''_n defines the maximum value of the stress intensity factor at the crack tip. In general, only the maximum value of the stress intensity factor is of interest. It means that φ can be defined when K_{III} reaches the maximum value. With the same method used to solve Eqs. (15a, 15b), we find the solutions as listed in Table II. The solutions can now be used to get $A_1(s)$, $A_2(s)$ and $p_2(x_0)$. The results may be used to calculate the stress intensity factor from stress distribution along the dislocation. The stress intensity factor at the crack tip from Eq. (10a) may now be defined as follows:

Table I $\alpha = 1$ and $\theta = \pi/2$

a_n	b_n
-0.178	-0.297
0	0
0.01	0.014
0	0
$-1.64\cdot10^{-4}$	$-2.22\cdot10^{-4}$
0	0

$$|K_{III}| = |\lim_{x\to 1}[2(x-1)]^{1/2}\sigma^*| \quad (19)$$

where $|K_{III}|$ is the stress intensity factor at the crack tip. At the regular points, the value of the stress intensity factor will be zero. For the singular points, the stress intensity factor is expressed as:

Table II $\alpha = 1, \theta = \pi/2, x_0 = 2, \mu = 0.3$

a_n	b_n
0.027	0.026
0.001	0.003
$-6.02\cdot10^{-4}$	$9.99\cdot10^{-4}$
$1.58\cdot10^{-4}$	$5.95\cdot10^{-4}$
$3.02\cdot10^{-4}$	$3.77\cdot10^{-4}$
$4\cdot10^{-4}$	$1.49\cdot10^{-4}$

$$|K_{III}| = \lim_{x\to 1}[2(x-1)]^{1/2}(-4/\pi)\{\sum_{n=0}^{N}[a_n \sin\omega t + b_n \cos\omega t \int_{-1}^{1} U_n(s)\frac{(1-s^2)^{1/2}}{(x-s)^2}ds] \quad (20)$$

According to [14], the integration in Eq. (20) is expressed as:

$$\int_{-1}^{1}\frac{(1-s^2)^{1/2}}{(x-s)^2}ds = -\frac{x\pi}{(x^2-1)^{1/2}}, \quad x \geq 1 \quad (21)$$

This leads to the following equation:

$$K_{III} = 4\{[\sum_{n=0}^{N}\frac{a_n}{n+1}]^2 + [\sum_{n=0}^{N}\frac{b_n}{n+1}]^2\}^{1/2}\cos(\omega t-\theta_n) = |K_{III}|\cos(\omega t-\theta_n) \quad (22)$$

where $\theta_n = \tan^{-1}\dfrac{\sum_{n=0}^{N} a_n (n+1)^{-1}}{\sum_{n=0}^{N} b_n (n+1)^{-1}}$ and K_{III} is the stress intensity factor of mode III.

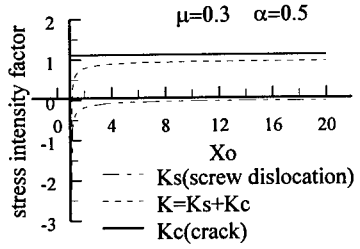


Figure 4 Effect of a screw dislocation on the SIF

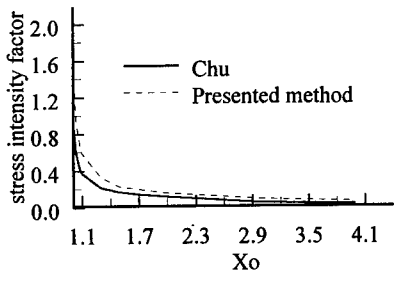


Figure 5 Comparison of dislocation and Chu model

The stress intensity factor K_{III} at the crack tip for a mode III surface crack is defined as follows:

$$K_{III} = K_S + K_C \quad (23)$$

where the first term, K_S, is due to stress field of the dislocation and the second term, K_C, is doe to the applied SH waves. Both of them derive from Eq. (22). Eq. (23) is plotted in Fig. 4 for arbitrary values of α and x_0. It is seen that the net stress intensity factor K_{III} can be reduced by a screw dislocation generated in the vicinity of the crack tip. It also can be shown that $K_{III} \approx 0$ when $x_0 \geq 20$. From Eqs. (23), we know the stress intensity factor can also be enhanced by the presence of the dislocation depending upon the relative sign of the applied SH waves and the Burgers vector of the dislocation. By holding the screw dislocation stationary, the strain energy release rate corresponding to the virtual displacement of the crack tip gives the crack extension force G_{III}. The G_{III} is defined as:

$$G_{III} = \frac{K_{III}^2}{2\mu} \quad (24)$$

The change of G_{III} as the surface crack propagation towards the screw dislocation is illustrated in Fig. 4. Without the applied SH waves, the dislocation stress fields tends to drag the surface crack to the dislocation where the strain energy can be relaxed. Under the applied SH waves, the surface crack can be repelled by the dislocation depending upon the direction of the applied SH waves and the Burgers vector the screw dislocation. There is a stable position x_0 where the crack extension force G_{III} is zero. In Fig. 5, the SIF curve for $\alpha = 0$ (static) is compared with that given in Chu [5]. In Fig. 6, it shows the dynamic stress intensity factors will increase at low frequency when the α increases and reach the maximum value. At relatively high frequencies, the stress intensity factor drops rapidly beyond the first maximum value and exhibits oscillations as α increases. To extend the study further to include the relation between the input angle θ and stress intensity factor shown in Fig. 6. It shows that the stress intensity factor increases with increasing the input angle θ at low frequencies. It is noted that the stable position x_0 will decrease when the input angle θ increases as shown in Fig. 7.

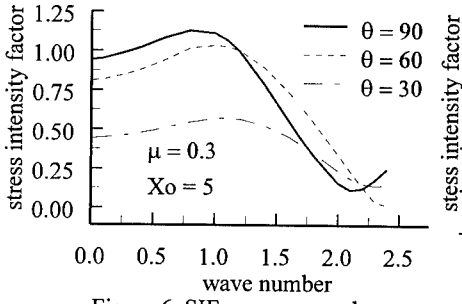


Figure 6 SIF vs wave number and incident angle

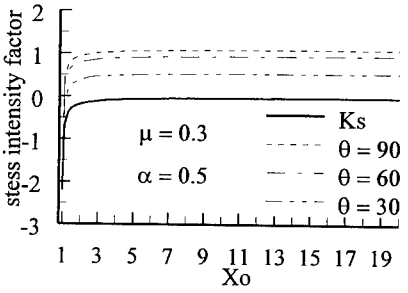


Figure 7 SIF vs x_0 and incident angle

THE INTERNAL CRACK: Our method can be extended for the internal crack near a free surface, using the mirrored dislocation method. Consider the configuration in Fig. 8, the crack of Fig. 1 is now moved in the x direction by b + 1. The stress released from the crack is equal to the anti-plane shear waves at $b \le x \le a$. Using the method of Shiue and Lee [15], we obtain the stress field S(x) of normalized length of a surface crack :

$$S(x) = \sigma_a \, x/(x^2-1)^{1/2} \tag{25}$$

Where σ_a is applied stress. Again, in order to satisfy the boundary condition along the crack surface, the S(x) must equal to the stress released from the screw dislocation (crack). Thus:

$$S(x) = -\int_{-1}^{1} D(s)\,\sigma_d(x-s)ds = -\int_{-1}^{1} D(s)\,B\alpha^2\{[J_0(\alpha|x-s|) + J_2(\alpha|x-s|)]\cos[\omega t+p(s)] + [Y_0(\alpha|x-s|)+$$

$$Y_2(\alpha|x-s|)]\sin[\omega t+p(s)]\}ds = -4\{[\sum_{n=0}^{N} \frac{a_n}{n+1}]^2 + [\sum_{n=0}^{N} \frac{b_n}{n+1}]^2\}^{1/2}\cos(\omega t-\theta_n)\, x/(x^2-1)^{1/2} \tag{26}$$

Hence the σ_a can be expressed in the following form:

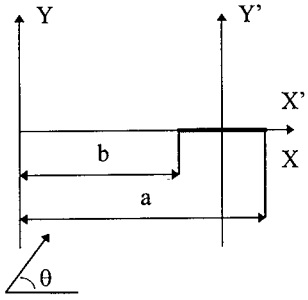


Figure 9 Internal crack

$$\sigma_a = -4\{[\sum_{n=0}^{N}\frac{a_n}{n+1}]^2 + [\sum_{n=0}^{N}\frac{b_n}{n+1}]^2\}^{1/2}\cos(\omega t-\theta_n) \quad (27)$$

Applying the method of Lee [16], the stress field S'(x) of normalized length of internal crack is:

$$S'(x) = \sigma_a [x^2-a^2E(\kappa)/G(\kappa)]/(x^2-a^2)^{1/2}(x^2-b^2)^{1/2} \quad (28)$$

where κ^2 equals to $(a^2-b^2)/a^2$. $G(\kappa)$ and $E(\kappa)$ are the first and second kinds of complete elliptic integrals as defined as:

$$E(\kappa) = \int_0^{\pi/2}\frac{d\beta}{\sqrt{1-\kappa^2\sin^2\beta}} \quad (29a)$$

$$G(\kappa) = \int_0^{\pi/2}\sqrt{1-\kappa^2\sin^2\beta}\,d\beta \quad (29b)$$

where $0 \le \beta \le 2\pi$. Substituting Eq. (27) into Eq. (28), the stress intensity factors at the right-hand and left-hand side crack tips are obtained as:

$$|H_{III}|^R = |\lim_{x\to a+}[2(x-a)]^{1/2}\tau_{yz}| = (\frac{a^2}{a^2-b^2})^{1/2}[1-E(\kappa)/G(\kappa)]a^{1/2}|K_{III}| \quad (30a)$$

$$|H_{III}|^L = |\lim_{x\to b-}[2(x-b)]^{1/2}\tau_{yz}| = (\frac{b^2}{a^2-b^2})^{1/2}[a^2E(\kappa)/b^2G(\kappa)-1]b^{1/2}|K_{III}| \quad (30b)$$

$|K_{III}|$ is derived from Eq. (22). In order to understand the effect of free surface on the stress intensity factor, the Fig. 9, 10 are plotted ($\theta = \pi/2$) to analyze the relation between the distance a and the stress intensity factors. In Fig. 9, the $|H_{III}|^R$ decreases with increasing b. When b approaches infinity, the $|H_{III}|^R$ will approach to $|K_{III}|$. Then crack near a free surface can be known as a crack embedded in an infinite solid. When we increase b, the $|H_{III}|^L$ will decrease and approaches $|K_{III}|$ while b approaches infinity. In Eq. (30b), the $|H_{III}|^L$ will become infinity when b close to zero (but not zero). It means an internal crack which is very close to a free surface will be easily extended to a surface crack. Therefore, there are two special cases worthy of mention. First, when b is equal to zero, the problem becomes that of a surface crack. Secondly, when b approaches infinity, then the problem is reduced to the case of an internal crack. The effect of interference between the input incident angle θ and the stress intensity factors also can be clearly seen in Fig. 11, 12. It is noted that results are in good agreement with that of Murakami [17]. It means that Eqs. (30a, 30b) not only satisfy the dynamic problems also agree with static's.

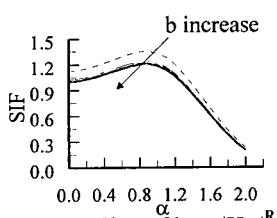


Figure 9 Effect of b on $|H_{III}|^R$

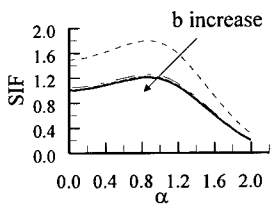


Figure 10 Effect of b on $|H_{III}|^L$

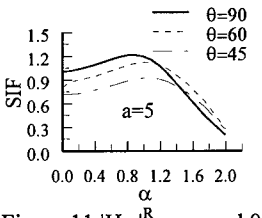


Figure 11 $|H_{III}|^R$ vs α and θ

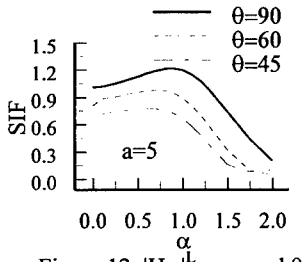


Figure 12 $|H_{III}|^L$ vs α and θ

CONCLUSION: The effects of dynamic loading for a mode III surface crack interacting with a screw dislocation are presented. The Chebyshev Polynomials, based on the stress boundary condition of the crack surface, are presented for obtaining the stress intensity factor of the crack tip. In absence of the applied SH waves, dislocations are always attracted towards the crack. When the SH waves are in the direction of driving the dislocation into the medium, a stable position is created at some distance from the crack tip beyond which the dislocations will be repelled by the crack. The stable position decreases with increasing input angle θ. At low frequencies, the stress intensity factor increases with increasing the input angle θ. The stress intensity factor can also be enhanced by the presence of the screw dislocation depending upon the relative sign of the Burgers vector of the dislocation.

REFERENCE :

[1]. Louat, N. P., *Proc. of the First Int. Conference on Frac.*, Japan, pp. 117-132 (1965).
[2]. Smith, E., *Acta Metall.* 14, pp. 556-557 (1966).
[3]. Rice, J. R. and Thomson, R., *Philos. Mag.* 29, pp. 73 (1974).
[4]. Shiue, S. T. and Lee, S., *Eng. Fract. Mech.* 22, pp. 1105 (1985).
[5]. Chu, S. N. G., "Elastic Interaction Between a Screw Dislocation and Surface Crack," *J. Appl. Phys*, Vol. 53, pp. 8678-8685 (1982).
[6]. Lee, S., "A New Analysis of Elastic Interaction Between a Surface Crack and Parallel Screw Dislocations," *Eng. Fracture Mechanics*, Vol. 22, pp. 429-435 (1985).
[7]. Li, J. C. M., *Dislocation Modeling of Physical Systems*, edited by Ashby, M. F. *et al.*, Pergamon, New York, pp. 498-518 (1981).
[8]. Warren, W. E., "Interaction of Dislocation with Surface Notches and Protrusions," *Int. J. Eng. Sci.*, Vol. 8, pp. 545 (1970).
[9]. Sih, G. C., "Some Elastodynamic Problems of Crack," *Int. J. Fracture Mechanics*, Vol. 4, pp. 51-68 (1968).
[10]. Eshelby, J. D., "Dislocation as a Cause of Mechanical Damping in Metals," *Proc. Royal Soc.*, London, A197, pp. 396-416 (1949).
[11]. Nabarro, F. R. N., "The interaction of screw dislocations and sound waves," *Proc. Royal Soc.* (London), A209, pp.278-290 (1951).
[12]. Loeber, J. F. and Sih, G. C., "Diffraction of Antiplane Shear Wave by a Finite Crack" *J. Acoust. Soc. Am.*, Vol. 44, pp. 90-98 (1968).
[13]. Rivlin, T. J., *The Chebyshev Polynomials*, Wiley Interscience, pp. 1-34 (1974).
[14]. Gradshteyn, I. S. and Ryzhik, I. M., *Table of Integrals Series*, pp. 833-836 (1965).
[15]. Juang, R. R. and Lee, S., " Elastic Interaction Between General Parallel Dislocations and a Surface Crack," *J. Appl. Phys.*, Vol. 59, pp. 3421-3429 (1986).
[16]. Shiue, S. T. and Lee, S., "Elastic Interaction Between Screw Dislocations and the Internal Crack Near a Free Surface", *Int. J. of Fracture*, Vol. 49, pp. 29-40 (1991).
[17]. Murakami, Y., *Stress Intensity Factors Handbook*, Pergamon Book Ltd., Vol. 2, pp. 169-170 (1987)

WAVELET TRANSFORMATION : THEORY AND ENGINEERING APPLICATIONS

D. Paul†, S. M. Pandit‡, and W. Hernandez†
† Monitoring Technology Corporation
Fairfax, VA 22031
‡ Michigan Technological University
Houghton, MI 49931

Abstract: The wavelet transformation (WT) overcomes the limitations of Fourier Transformation (FT). In comparison to FT, where only the global information is available, WT provides the local information by projecting the signal into the time-frequency plane. This makes WT a powerful signal analysis tool; but the effective use of wavelet is impeded due to the complex mathematical framework. This paper provides an easy-to-understand reading on the theory of wavelet, avoiding mathematical complexity. The powerful capabilities of wavelet coefficients are explored. Three examples demonstrate the application of WT in the areas of vibration analysis and manufacturing fault diagnosis. Therefore, this paper will be very interesting to engineers to understand the theory, implementation, and application of WT.

Key Words: Diagnostics; signal processing; vibration analysis; wavelets

INTRODUCTION: Wavelet Transformation (WT) is the most recent state-of-the-art technique which decomposes signals into time-frequency plane. Usually, the signal is obtained in the time domain. Fourier Transformation (FT) decomposes the signal into the frequency domain; thus the various frequency components of the signal are obtained. WT adds one more dimension, the time information, to the decomposition. As a result, not only the frequency components, but also their duration can be known. This supplementary time dimension makes WT very powerful.
The idea of wavelet transformation is not new; a similar transformation was initiated by Haar in 1910. The wavelets first came into the limelight when Grossmann and Morlet defined the continuous wavelet transformation in 1984 [6]. The search was continued for a class of orthogonal basis functions that have a finite duration both in time- and frequency-domain. A major breakthrough in the history of wavelets occurred, when I. Daubechies constructed such functions in 1988 [3].

FOURIER TRANSFORMATION TO WAVELET TRANSFORMATION: The Fourier transformation (FT) of a signal $x(t)$ is defined by:

$$\hat{x}(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt =< x(t), f_\omega(t) > \qquad (1)$$

where $\hat{x}(\omega)$ is the Fourier transform of the signal $x(t)$ and the symbol $<,>$ indicates the inner-product. The function $f_\omega(t) = e^{-i\omega t}$, which is sinusoidal, is the basis function of frequency ω. The class of basis functions is generated by varying the frequency

ω. For continuous case, the frequency ω varies continuously. However, in practice, we always deal with the discrete case; where the basis functions are generated [2] by $f_n(t) = e^{-i\omega_0 nt}$. The parameter $n = 1, 2, \cdots$ is the frequency index and ω_0 is the fundamental frequency. The index n is known as the *dilation* parameter and the basis functions are generated by *dilating* the original function $f(t) = e^{-i\omega_0 t}$.
The sinusoidal functions $e^{-i\omega_0 nt}$ have two important properties: (1) They are orthogonal to each other and (2) they have small duration in frequency domain. The first property helps FT to preserve the signal energy (the energy of $x(t)$ is defined by $I\!E_x = \int |x(t)|^2 dt$), that is $I\!E_x = I\!E_{\hat{x}}$. This helps the straight forward reconstruction of the signal:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{x}(\omega) e^{i\omega t} d\omega \tag{2}$$

The term *resolution* is widely used with wavelet transformation. The time-resolution σ_f of the signal $f(t)$ indicates that two short-time signals can be discriminated if they are σ_f apart. Similarly, the frequency resolution $\sigma_{\hat{f}}$ of $\hat{f}(\omega)$ indicates that two frequency components can be discriminated if they are $\sigma_{\hat{f}}$ apart [11]. It is clear that the basis function of FT has the finest frequency resolution; thus it should separate any two frequency components. However, due to the infinite time-duration (hence, infinite time-resolution), FT is unable to separate two signals of same frequency but different time durations.
The time-resolution problem in FT is overcome in WT by introducing an additional time index (similar to the frequency index n) in the basis function. Continuous Wavelet Transformation (CWT) of a signal $x(t)$ is defined by [5]:

$$W_{ab}^x = \int_{-\infty}^{\infty} x(t) \psi_{a,b}(t) dt = < x, \psi_{a,b} > \tag{3}$$

The dilation parameter a represents the frequency; the parameter b indicates the translation in time. The basis function $\psi_{ab}(t)$ (known as daughter wavelet) is the dilated and translated version of the mother wavelet $\psi(t)$ such that

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi \left(\frac{t-b}{a} \right)$$

$\frac{1}{\sqrt{|a|}}$ is the normalization factor so that the energy of daughter wavelets remains the same as that of the mother wavelets. W_{ab}^x is also known as the wavelet coefficient.

Basis function for WT: As can be seen, like FT, WT is also the inner-product of $x(t)$ and $\psi_{a,b}(t)$. However, note that $\psi_{a,b}(t)$ should have short time and frequency duration in order to get acceptable resolution in both domains. Therefore, in addition to the properties of FT basis functions (orthogonality and fine frequency resolution), the WT basis functions need short time duration too. For a long time, researchers were searching for functions having above three properties.
However, in the case of CWT, one property (orthogonality) can be sacrificed, because, it has been shown that CWT preserve energy even for non-orthogonal basis [11].

Later, I. Daubechies constructed a set of orthogonal functions that have finite time- and frequency-resolution.
So, unlike the sinusoidal function, $\psi(t)$ should die out (decaying function) eventually to provide finite time-resolution. Also, we expect $\psi(t)$ to be sufficiently smooth, because any discontinuity in one domain is spread over the entire other domain. These conditions imply that $\psi(t)$ should oscillate in time like a short wave [11]. Thus, an essential condition for any function $\psi(t)$ to be a wavelet basis is:

$$\int \psi(t)dt = 0, \tag{4}$$

which is equivalent to $\hat{\psi}(0) = 0$ ($\hat{\psi}(\omega)$ is the FT of $\psi(t)$).
Therefore, $\psi(t)$ has short duration in both domains. To cover the entire frequency axis/domain, we need to dilate $\psi(t)$ by the changing a (similar to n in FT). Unlike the sinusoidal basis of FT, (which readily covers the entire time-axis), the $\psi(t)$ also needs to be translated using the parameter b to cover the entire time-axis.

Wavelet Transformation - a multiresolution technique: The time- and frequency-resolution are complementary, and their relation is bounded by the *uncertainty principle:*

$$\sigma_x \sigma_{\hat{x}} \geq \frac{1}{2}$$

Thus, we cannot choose arbitrarily small resolution in both domains. If we go for a finer resolution in one domain, the other domain's resolution will become coarser. The resolutions for mother and daughter wavelets are not the same. Let the center (the point of symmetry) and resolution for $\psi(t)$ be μ_ψ and σ_ψ, in time domain and $\mu_{\hat{\psi}}$ and $\sigma_{\hat{\psi}}$, in frequency domain. Using the basic properties of FT [1], it can be shown that corresponding parameters for $\Psi(t) = \psi_{ab}(t)$ are:

$$\mu_\Psi = a\mu_\psi + b, \sigma_\Psi = a\sigma_\psi, \mu_{\hat{\Psi}} = \frac{\mu_{\hat{\psi}}}{a}, \sigma_{\hat{\Psi}} = \frac{\sigma_{\hat{\psi}}}{a} \tag{5}$$

Note that the resolution depends on the dilation parameter a. Furthermore, the ratio $\mu_{\hat{\psi}}$ (center) to $\sigma_{\hat{\psi}}$ (resolution):

$$\frac{\mu_{\hat{\Psi}}}{\sigma_{\hat{\Psi}}} = \frac{\mu_{\hat{\psi}}}{\sigma_{\hat{\psi}}} \tag{6}$$

is constant (does not depend on the frequency). This is an important property of wavelet transformation, where time- and frequency-resolution depend on the scale a. Eq. 5 shows that for the lower frequency (larger a), WT has finer frequency resolution with coarser time resolution and for higher frequency, the opposite occurs. Due to this variable resolution, WT is known as a *multi-resolution* technique.

DISCRETE WAVELET TRANSFORMATION: In CWT, the dilation parameter a and the translation parameter b vary continuously, whereas in discrete wavelet transformation, they take only discrete values. Discretization of a and b can be done as [4]:

$$a = a_0^{-m}, b = nb_0a_0^{-m}, \text{for } a_0 > 0, b_0 > 0 \tag{7}$$

With $a_0 = 2, b_0 = 1$ (for computational efficiency), the daughter wavelet becomes:

$$\psi_{mn}(t) = 2^{\frac{m}{2}}\psi(2^m t - n) \tag{8}$$

where m indicates the scale (frequency index) and n is the time index indicating location. However, due to the discretization, an important feature of CWT is lost; discrete wavelet preserves energy only when $\psi_{mn}(t)$ are orthogonal. This adds a severe restriction for choosing basis functions.
Mainly, there are two forms of discrete wavelets: Fast Wavelet Transformation (FWT) and Dyadic Wavelet Transformation (DWT). The next sub-section shows a step-by-step derivation of FWT from the classical definition of WT.

Fast Wavelet Transformation: To explain FWT, we need to use some definitions and properties of spaces. Let consider a sequence of nested subspace V_m such that a higher dimensional space V_{m+1} includes lower dimensional space V_m. For example, the well known 3-D space generated by x-, y-, and z-axes includes the lower 2-D plane generated by x- and y-axes only.
A larger dimensional space V_{m+1} can be decomposed into orthogonal components V_m and W_m. Again for example, the 3-D space can be considered as a sum of 2 orthogonal spaces: the 2-D space generated by $x - y$ axes and the 1-D space generated by z-axis. Note that any function either in V_m or in W_m are also in V_{m+1}.
From the linear algebra, we know that each space is spanned by a set of independent vectors (functions); like vectors $x = \{1, 0, 0\}, y = \{0, 1, 0\}$, and $z = \{0, 0, 1\}$ that span 3-D space. Similarly, let the reference subspaces V_0 and W_0 be spanned by the scaling function $\phi(t) = \phi_{0,n}(t)$ and wavelet function $\psi(t)$, respectively. An important property of V_m subspaces is that if V_0 is generated by $\phi_{0,n}(t)$, then the space V_1 is generated by the dilated and translated version of $\phi(t)$: $\phi_{1,n}(t) = 2^{\frac{1}{2}}\phi(2t - n)$.
Both V_0 and W_0 are in V_1. Therefore, both the scaling and wavelet functions can be expressed as a linear combination of $\phi_{1,n}(t)$:

$$\phi(t) = \sum_n p_n \phi(2t - n), \qquad \psi(t) = \sum_n q_n \phi(2t - n) \tag{9}$$

Suppose a function $x(t)$ is in V_m so that:

$$x(t) = \sum_n c_{m,n}\phi_{m,n}(t) \tag{10}$$

$$\text{where} \quad c_{m,n} = < x, \phi_{m,n} > \tag{11}$$

Since V_m consists of V_{m-1} and W_{m-1}, the function $x(t) \in V_m$ can be decomposed as ($x_{m-1} \in V_{m-1}$ and $y_{m-1} \in W_{m-1}$):

$$x(t) = x_{m-1}(t) + y_{m-1}(t) \tag{12}$$

$$= \sum_l c_{m-1,l}\phi_{m-1,l}(t) + \sum_l d_{m-1,l}\psi_{m-1,l}(t) \tag{13}$$

where $d_{m-1,l} =< x, \psi_{m-1,l} >$ is known as the wavelet coefficient at scale $(m-1)$ and time l. Further,

$$c_{m-1,l} = \int x(t)\phi_{m-1,l}(t)dt$$

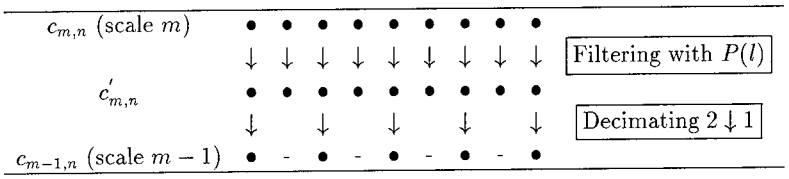


Figure 1: Coefficients at various levels

$$= \sum c_{m,n} \int \phi_{m,n}(t)\phi_{m-1,l}(t)dt \tag{14}$$

For dyadic case (Eq. 8):

$$\phi_{m-1,l} = 2^{\frac{m-1}{2}}\phi(2^{m-1}t - l)$$

Let, $2^{m-1}t - l = s$, thus, $dt = 2^{1-m}ds$. Substituting these values,

$$\begin{aligned} \phi_{m,n}(t) &= 2^{\frac{m}{2}}\phi(2^m t - n) \\ &= 2^{\frac{m}{2}}\phi(2l + 2s - n) \end{aligned} \tag{15}$$

Therefore,

$$\begin{aligned} c_{m-1,l} &= \sum_n c_{m,n} 2^{\frac{m-1}{2}} 2^{\frac{m}{2}} 2^{1-m} \int \phi(2l + 2s - n)\phi(s)ds \\ &= \sum_n c_{m,n}\sqrt{2} \int \phi(2l + 2s - n)\phi(s)ds \end{aligned}$$

Let $\sqrt{2}\int \phi(2l + 2s - n)\phi(s)ds = P_{2l-n}$, so that $c_{m-1,l} = \sum_n c_{m,n}P_{2l-n}$. Using, $c'_{m,l} = \sum_n c_{m,n}P_{l-n} = c_{m,l} \star P_l$ ($\star$ denotes convolution),

$$c_{m-1,l} = c'_{m,2l} \tag{16}$$

Therefore, $c'_{m,l}$ can be viewed as a convolution between $c_{m,l}$ and P_l. Convolution and filtering are the same operations; that is, $c'_{m,l}$ is the filtered version of $c_{m,n}$ using P_l as the filter coefficients. Since $c_{m-1,l} = c'_{m,2l}$, $c_{m-1,l}$ consists of every other point of $c'_{m,2l}$. Thus, if there are M coefficients at scale m, the number of coefficients at scale $(m-1)$ is $\frac{M}{2}$. This phenomenon is illustrated in Fig. 1.
With similar substitution, it can be shown that:

$$d_{m-1,l} =< x, \psi_{m-1,l} >= \sum_n c_{m,n}Q_{2l-n} \tag{17}$$

where $Q_{2l-n} = \sqrt{2}\int \phi(2l + 2s - n)\psi(s)ds$.
Similar to $c_{m-1,l}$, the wavelet coefficients at scale $(m-1)$ can be found by filtering $c_{m,n}$ using Q_l as the filter coefficients and discarding every alternate points. P_l are

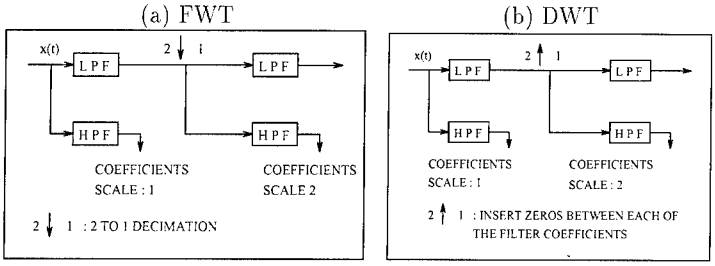


Figure 2: Discrete Wavelet Transformation

the coefficients of the low pass filter (LPF), whereas Q_l represents a high pass filter (HPF). The FWT is schematically shown in Fig. 2(a).

Dyadic Wavelet Transformation: The Eq. 3 can be modified by substituting $t = b - t'$ [7]:

$$W_{ab}^{x} = \int_{-\infty}^{\infty} x(b-t')f_a(t')dt' = x(b) \star f_a(b) \tag{18}$$

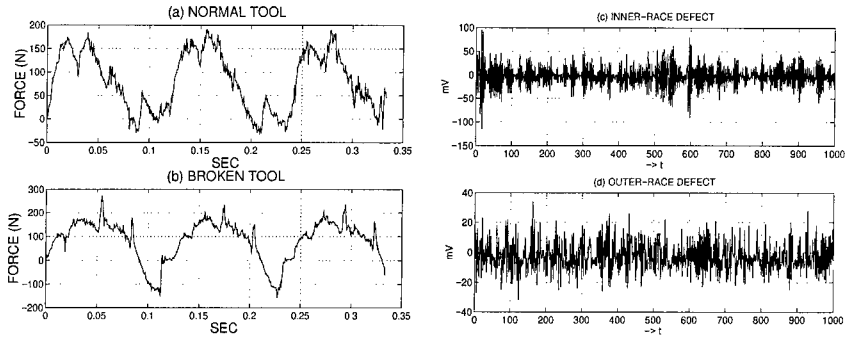
where $f_a(t') = -\frac{1}{\sqrt{|a|}}\psi(-\frac{t'}{a})$. Thus, the wavelet transformation can also be defined as a convolution.
The dyadic wavelet transformation (DWT) is defined as a convolution of the signal, $x(t)$, and the basis function $h_{2^j}(t)$ [8, 9]:

$$W_{2^j}^{x}(t) \;=\; x(t) \star h_{2^j}(t) \tag{19}$$

where $h_{2^j}(t) = \frac{1}{2^j}h(\frac{t}{2^j})$. $W_{2^j}^{x}(t)$ is the DWT coefficient at scale j. The discretization of the parameter a takes place in the dyadic scale, $a = 2^j$. However, the coefficients are calculated in each location. Thus, DWT is over-complete and redundant and it does not preserve the energy. DWT is schematically shown in Fig. 2(b).

WAVELET COEFFICIENTS : POST-PROCESSING WITH APPLICATIONS: In this section, the capabilities of wavelet coefficients are explored. Various detection algorithms are derived by post-processing the coefficients. The important system parameters are related with wavelet coefficients to identify a variety of manufacturing defects. The algorithms are tested through experimental signals (data).

Example 1 : Tool breakage detection in end-milling: In an end milling operation, each time that a tooth of the tool starts cutting, the impact between the tooth and the work-piece generates an impulse. This impulse propagates from the impact source to the sensor location, exciting the underlying system. Thus, the mode produced by the impact is the impulse response of the system. Due to the nature of the system, the signal is essentially a damped sine wave.
The impact is generated at a fixed interval corresponding to the tooth period (T_r). Therefore, the signal from the normal tool cutting operation consists of a series of



End-Milling : Force-data Bearing : Vibration-data

Figure 3: Force and vibration data

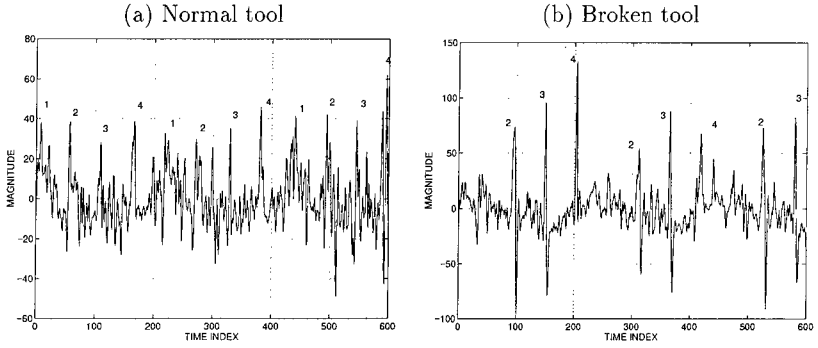
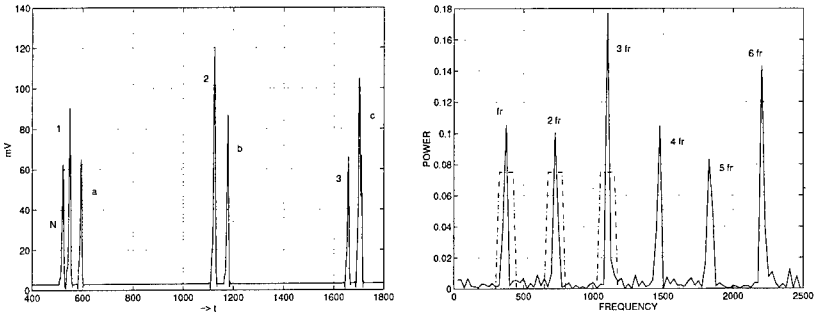


Figure 4: DWT coefficients

damped sine waves spaced at the tooth period. In the case of the broken tool, the broken or worn-out tooth removes little or no material. Thus, there are very small (if any) impacts produced by this tooth. Hence, the impact mode of the broken tool has missing peaks. The number of missing peaks per revolution equals to the number of broken teeth. As DWT coefficients provides the time-variation of the signal at a particular scale, they can be used to differentiate the normal- and broken-tool modes.

In this example, two force data sets (experimental) obtained from an end-milling operation are considered. The experiment was conducted using two, four-fluted tools: one tool is normal [Fig. 3(a)], the other tool has one tooth ground down to eliminate it's cutting capability [Fig. 3(b)].

The dyadic wavelet coefficients at scale 2 for normal and broken tools are shown in Fig. 4. In the case of normal tool, peaks at regular interval of tooth period ($t = 54$) are present. (Also, some noise peaks are present at the same time; but they do not repeat at any particular intervals.) This confirms that each tooth is cutting material; thus



(a) Envelope Signal (b) Diagnostic Spectrum

Figure 5: Detection of Inner- and Outer-race Defects

producing impacts at 4 instances for one shaft revolution (time index for 1 revolution = 216) [marked by 1,2,3, and 4 in Fig. 4(a)]. The situation is different in the case of a broken tool [Fig. 4(b)]. In this case only 3 (2nd, 3rd, and 4th) peaks are prominent during one shaft revolution. The first peak is missing from each cycle. So, there is no impact between the first tooth and the work-piece.

Example 2 : Envelope signal: The signal component $X_a(t)$ for any scale a can be reconstructed using FWT coefficients of that scale only.

$$X_a(t) = \sum_b W_{ab}^x \psi_{ab}(t) \tag{20}$$

which can be considered as zoomed signal in time domain. Due to the filtering of other components, the noise level in $X_a(t)$ is expected to be much less than the original signal. The noise level can be further reduced by discarding the coefficients of small magnitude before reconstruction.

The signal for this example is taken from a bearing assembly consisting of 4 bearings [Fig. 3(c)]. One of the bearing has a defect in the inner-race. The signal produced by the defective bearing is buried under the noise produced by other bearings and it is difficult to identify the defect from the signal.

The FWT coefficients have been calculated using the Daubechies (6th order) wavelets. The component $X_a(t)$ has been reconstructed using the coefficients at scale 2 that are greater than $0.4Max$ (Max is the maximum value of the wavelet coefficients at scale 2). In the next step, an envelope of $X_a(t)$ is generated using Hilbert Transformation, which is shown in Fig. 5(a). The envelope signal shows one peak per revolution (time index for one revolution = 570) as numbered by 1, 2, and 3. Each of the peak has a side band placed at the defect period T_r = 52 corresponding to inner-race defect (marked by a, b, and c). A noise spike (marked by N) can also be seen in the plot. The envelope signal, generated without any system information, clearly indicates the inner-race defect.

Example 3 : Diagnostic spectrum for discrete defects: To get conclusive

results for defect diagnosis in the time domain, the defect spikes should be present (in the processed data) consistently at the defect repetition period T_r. In the frequency domain, one peak at defect frequency f_r would be enough to confirm the presence of defect. Hence, the spectrum would be helpful for automated monitoring.

In this last example, wavelet coefficients are post-processed to get a diagnostic spectrum, where the discrete defects can be identified at the early stage. The abrupt changes in the signal produced by the discrete defects can be captured by *extrema* (the local maxima and minima) of DWT coefficients [8]. However, extrema identify defects when it is significantly large. To get an early indication, further processing of extrema is needed.

Fig. 3(d) shows vibration signals from an bearing assembly, where one of the bearings has a spall on the outer-race. The signals produced by the defect are buried under the noise. To extract the defect spikes (the interval between spikes is T_r=70) from the noise, the autocorrelation of the DWT coefficients have been calculated [10].

For further reduction of noise, we need to go to the frequency domain. Usually, the noise is discarded by choosing a threshold value [10]. In practice, not only it is difficult to choose the threshold value, but also the threshold value affect the performance. A noise reduction scheme is derived avoiding the choice of threshold value.

The data was collected for 2 shaft revolutions ($t = 2 * 570$). Defect repetition period being $T_r = 70$, only 8 ($570/70 \approx 8$) spikes are expected per revolution. Thus, there should be 16 spikes for 2 revolutions. A few-lines of code using *Matlab* is written to pick first 16 peaks from the autocorrelation data (2 revolutions). In the ideal case, the code will capture all 16 peaks produced by the outer-race defect; but in the real situation, some noise peaks may be sensed instead of defect peaks. For this example, 14 defect peaks (missing at $t = 910$ and 980) along with 2 noise peaks (for $t = 25$ and 580) are picked up.

In the next step a binary vector M (consists of 1 or 0) is created using the sensed peaks:

$$M = [m_j] = \begin{cases} 1, & \text{at peaks' locations} \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

The vector, M, can be divided into two components: $M = M^d + M^n$. The M^d component has 1's at the impulse locations, and all of the other elements of this vector are zero. Most of the elements of the M^n component are zero, except few 1's that are due to noise spikes. In the present case, M_d has 14 numbers of 1's mostly $t = 70$ apart; whereas M_n has only 2 numbers of 1's at $t = 25$ and 580. In the spectrum of the vector M, the energy from the M^d component is concentrated at the defect frequency $f_r = 370$ Hz and its harmonics, whereas the energy from the M^n component is distributed over the entire frequency axis. As a result, the spectrum plot of the vector M has very distinct peaks at the locations of the defect frequency and its harmonics as shown in Fig. 5(b).

To make the detection scheme suitable for automated monitoring, an alarm level can be set. If the peaks at f_r exceeds the limit, the alarm will sound. In this example, an alarm is set at 0.075. This value is obtained from the spectrum of the ideal binary vector M, where all 16 defect spikes are captured (no noise spikes). The computer

generated rectangular boxes are overlayed in the spectrum (as shown by the dotted lines in Fig. 5(b)) over the defect frequency for monitoring the defect. Thus, the maintenance staff can very easily monitor the bearing health.

CONCLUSION: This article describes the theory, implementation, and applications of wavelet transformation. The usefulness of WT has been demonstrated through various practical examples. The ability to extract important features from the noisy signal gives credibility to WT. It can be concluded that WT will become a very powerful tool in the area of vibration/signal analysis in the near future. This paper will help users to understand the structure of WT and implement it in the areas of analysis and signal processing.

References

[1] A. N. Akansu and R. A. Haddad. *Multiresolution Signal Decomposition Transforms, Subbands, and Wavelets.* Academic Press, Inc., 1992.

[2] C. K. Chui. *An Introduction to Wavelets.* Academic Press, Inc., 1992.

[3] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communication on Pure and Applied Mathmatics*, 1988.

[4] I. Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, 1992.

[5] A. Grossmann, R. Kronland, and J. Morlet. Reading and understanding continuous wavelet transforms. *Wavelets (Editor: J.M. Combes and A. Grossmann)*, 1992.

[6] A. Grossmann and J. Morlet. Decomposition of hardy function into square integrable wavelets of constant shape. *SIAM Journal of Math. Anal.*, 15(4):723–736, 1984.

[7] C. J. Li and J. Ma. Bearing localized defect detection through wavelet decomposition of vibrations. *Sensors and Signal Processing for Manufacturing, WAM, ASME*, PED-Vol. 55:187–196, 1992. Edited by: S. Y. Liang and C. L. Wu.

[8] S. Mallat and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Transactions of Information Theory*, March 1992.

[9] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 1992.

[10] D. Paul. Detection of change in processes using wavelets. In *IEEE-SP Symposium on Time-Frequency and Time-Scale Analysis*, pages 174–177, 1994.

[11] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, pages 14–38, October 1991.

FAILURE ASSESSMENT OF CYLINDRICAL STRUCTURES SUBJECTED TO INTERNAL IMPULSIVE SPOT LOADING

M. Moatamedi, B.C.R. Ewan, J.L. Wearing

Department of Mechanical and Process Engineering
The University of Sheffield
Mappin Street, Sheffield S1 3JD, UK

Abstract: A transient analysis of a cylindrical shell structure under internal impulsive spot loading has been carried out using ANSYS finite element code. The spot loading has been chosen to be above the cylinder design pressure, and a number of applicable failure criteria namely Von Mises failure criterion and a specific dynamic criterion have been examined to investigate the failure of the structure under the specified conditions. The transient stress of the critical point in the analysis is presented to determine the time at which the maximum stress occurs. A critical curve is introduced to determine the critical internal impulsive spot loading and its duration for cylindrical structures, and the relations between impulse, pulse duration and the natural frequency of the structure are explored. It is indicated that the impulsive load on cylindrical structures may exceed the design pressure without structural failure. All analyses suggest that the design criteria for structures under dynamic loading are more flexible than those under static loading in which no freedoms in deviation of any simple yield criterion exist.

Key words: Design, failure, shell structures, spot loading, transient, yield criteria

INTRODUCTION: Cylindrical shell structures have a wide range of practical applications, some of which are subject to short duration transient loading. A clear understanding of the transient response and proper assessment of the failure of these structures subjected to impulsive loading require detailed analyses and experimental work. The transient analysis of a cylindrical shell structure subjected to internal impulsive spot loading using the ANSYS finite element code is addressed in this paper.

The design of vessels subjected to asymmetric pulse loading, particularly impulsive spot loading, poses problems totally different from those under axisymmetric loading. Some attempts have been made to investigate the failure and response of cylindrical structures under axisymmetric external pulse loading such as Pegg [1]. Molyneaux et al. [2] also studied the dynamic response of circular cylindrical shells subjected to axisymmetric internal pulse loading caused by detonating an explosive charge.

Lindberg [3] and Abrahamson and Goodier [4] have respectively proposed the elastic and plastic theories for the response of cylindrical shell structures under axisymmetric pulse

loading. Some numerical methods such as the analysis discussed by Oslon [5] have also been introduced for the elasto-plastic response of cylindrical shells under axisymmetric pulse loading since there is no general analytical solution for elasto-plastic behaviour of such structures subjected to pulse loading.

The study of Anderson and Lindberg [6] is on the pulse buckling of cylindrical structures subjected to impulsive lateral pressures and the critical pulse load for this condition has been introduced by Abrahamson and Lindberg [7]. A numerical study has also been carried out to investigate the effects of pulse and its duration on such conditions [8]. Although those efforts are on the response of cylindrical shells under asymmetric pulse loading, for many applications, failure and response of such structures to spot transient loading, need to be investigated. Although the response of cylindrical tanks to external impulsive spot loading has been studied by Schwer et al. [9], to the knowledge of the authors to the date, there is no work available on the response of cylindrical shell structures subjected to internal impulsive spot loading.

NUMERICAL ANALYSIS: The analyses have been performed, to predict the elasto-plastic response and failure of a cylindrical shell structure, using ANSYS code. ANSYS is a general finite element code with a wide range of capabilities including nonlinear transient analysis which enables the prediction of the response of the cylindrical structure under internal impulsive spot loading. It also allows the investigation of failure of the structure considering different defined criteria. In order to verify the accuracy of the finite element analysis for the prediction of the response of cylindrical structures to impulsive loading, the reader is referred to another study done by the authors [8]. The results of the mentioned study were compared to experimental results.

A 10-metre long circular cylindrical structure having a diameter of 1 metre made of steel, subjected to internal impulsive spot loading is analysed here. The spot, which has a diameter of 0.5 metre is located in one side on the middle of the cylinder. The cylinder was designed for 1 Mpa internal static pressure using ASME boiler and pressure vessel code [10]. In order to model the structure in ANSYS, 1512 shell elements with plasticity, large deflection, and large strain capabilities [11] are used as is illustrated in Fig. 1. Both ends of the cylinder are assumed to be clamped.

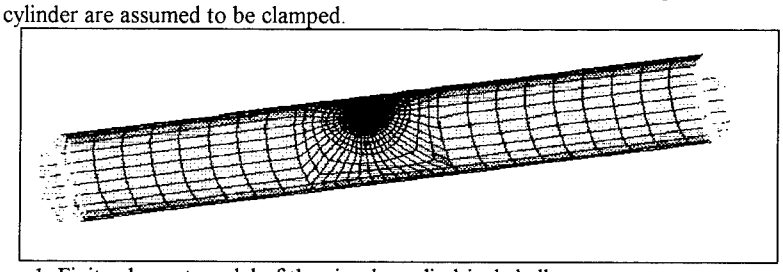


Figure 1. Finite element model of the circular cylindrical shell structure

Rectangular type of impulsive loading is used for all analyses so that the load suddenly increases to the maximum value and then remains for the limited duration and finally returns to zero.

The shell under impulsive loading is expected to have large deformations and thus large displacements and rotations which were taken into account by invoking the nonlinear strain-displacement relationships where higher-order derivatives of displacements and rotations are included. Nonlinearity of the material is assumed to follow the bilinear isotropic hardening. This option uses the Von Mises yield criterion to assess the failure of the structure under the specified loading conditions. Damping is also taken into account to increase the accuracy of the final deflection of the structure in order to obtain the realistic results.

The modal analysis of the cylinder was first undertaken to determine the natural frequency of the structure. This enables the calculation of the proper integration time step for the transient analyses. Integration time step was then calculated by the ANSYS recommended '*1/20f*' formula where '*f*' is the natural frequency of the structure [11]. A series of analyses were then carried out to observe the response of the structure to internal impulsive spot loading with the value of 4 Mpa and 10-50 msec duration. The failure of the structure, for each case, was checked by applying Von Mises criterion to determine the failure point of the structure.

RESULTS AND DISCUSSION: The fundamental frequency of the structure was determined by modal analysis to be 66.7 Hz (Period: 15 msec) which leads to the calculation of the integration time step '*1/20f*' for the analysis which equals 0.75 msec. The transient response of the centre of the spot, namely displacement and equivalent stress, obtained from the nonlinear transient analysis with 4 Mpa pressure and 10 msec duration, are shown in Fig. 2 and Fig. 3, respectively.

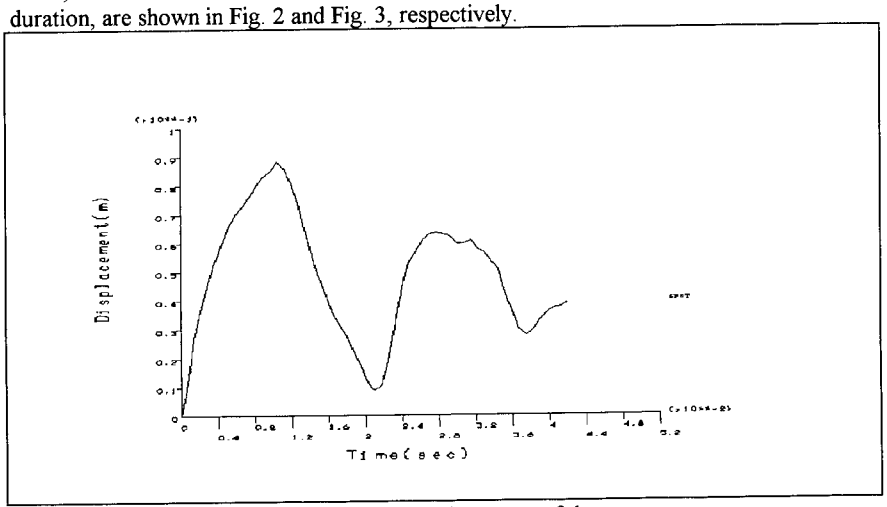


Figure 2. Displacement-time relationship for the centre of the spot

As can be observed in Fig. 2, the displacement of the point increases to a maximum value and then damps to a final deflection which is not zero. A combination of the structural modes are excited in dynamic response and therefore the natural frequency of the structure plays an important role in the analyses which is indicated as the structural response. If the pulse duration is very short compared to the structural response (the structural response is defined as the period of the structure which equals *'1/f'*), the response of the structure is perfectly plastic but a long duration results in an elastic response. For the present analysis, the duration is very close to the period of the structure so that the response is consequently elasto-plastic. The above explanation is correct when the structure is under the equal pulse in all mentioned cases.

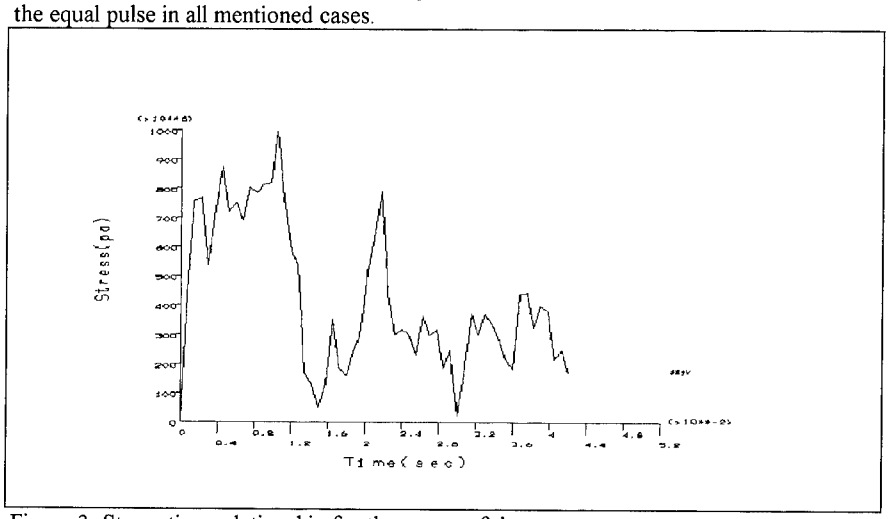


Figure 3. Stress-time relationship for the centre of the spot

The maximum stress (Fig. 3) in the structure reaches a value which is higher than the yield stress however the structure does not fail under this condition. The reason is that the pulse duration is too short to cause the failure so that the load is not categorised in static loading. It expresses that the structure can endure stresses higher than yield stress for less than a critical duration, in other words the dynamic yield criteria are different form the static ones.

In order to observe the effects of the duration of pulse loading on the response of the shell, a series of analyses on the model with different pulse duration have been carried out to indicate the relationship between the pulse loading duration with time at which the maximum displacement and stress occur. It was found that the maximum displacement of the centre of the spot, which is the critical point of analyses, occurs at 10.6 msec regardless of the pulse duration as long as the response is in elasto-plastic region. Therefore, the maximum displacement, in elasto-plastic response, does not depend on the pulse duration, however the final deflection which occurs in the plastic region depends on

the pulse duration which is due to the stress wave through the structure. Consequently, the role of the natural frequency of the structure is explored which basically provides the information on the structural response to indicate whether the response is elastic, plastic or elasto-plastic which is due to the pulse duration.

Von Mises failure criterion and also a dynamic criterion introduced by Schwer et al. [9] were applied to assess the structural failure for the present case. However the nonlinear transient analyses of the shell structure under internal transient spot loading with different values for over pressure and duration can produce a critical curve which indicates a safe region for cylindrical structures under impulsive spot loading without structural failure. The generation of the critical curve for the present case is experimentally under investigation by the authors.

CONCLUSIONS: The nonlinear transient analysis of a cylindrical structure due to an internal impulsive spot loading was carried out numerically using ANSYS finite element code. The response of the cylindrical vessel was found to be elasto-plastic since the pulse duration was close to the period of the structure. Von Mises and a dynamic failure criteria were applied to investigate the failure of the structure. According to the analysis carried out by ANSYS code, it was found that a dynamic pressure greater than the design pressure can be applied to the structure without failure if the impulse duration is less than a critical value. The critical value must be determined for any structure under specific conditions. It was also observed that the maximum displacement of the structure occurs at a certain time regardless the pulse duration as long as the structural response is elasto-plastic. It can be concluded that the design criteria for structures under dynamic loading are more flexible since the dynamic yield criteria are different from static ones.

REFERENCES:

[1] Pegg, N.G., *"A numerical study of dynamic pulse buckling of cylindrical structures"*, Marine structures 7, 189-212, 1994

[2] Molyneaux, T.C.K.; Li, Long-Yuan; Firth, N., *"Impact response of circular cylindrical shells under explosive loading"*, Advances in Engineering Software 18, 7-13, 1993

[3] Lindberg, H.E, *"Buckling of a very thin cylindrical shell due to an explosive pressure"*, Journal of applied mechanics 31, 261-272, 1964

[4] Abrahamson, G.R.; Goodier, J.N.; *"Dynamic plastic flow buckling of a cylindrical shell from uniform radial impulse"*, Forth US national congress of applied mechanics 2, 1962

[5] Oslon, M.D., *"Efficient modelling of blast loaded plate and cylindrical shell structures"*, Computers and structures, v40, n5, 1139-1149, 1991

[6] Anderson, D.L.; Lindberg, H.E., *"Dynamic pulse buckling of cylindrical shells under transient lateral pressures"*, AIAA Journal, v6, n4, 589-598, 1968

[7] Abrahamson, G.R.; Lindberg, H.E., *"Peak load-impulse characterisation of critical pulse loads in structural dynamics"*, Dynamic response of structures, Proceeding of a symposium held at Stanford university, California, 1971

[8] Moatamedi, M.; Ewan, B.C.R.; Wearing, J.L., *"Effects of explosions on the integrity of process plant structures"*, Process Plant Safety Symposium, AIChe, Houston, April 1996

[9] Schwer, L.; Holmes, B.S.; Kirkpatrick, S.W., *"Response and failure of metal tank from impulsive spot loading: experiments and calculations"*, International journal of solids and structures, v24, n8, 817-833, 1988

[10] American Society of Mechanical Engineers (ASME), *"Article D-3, Shells of revolution under external pressure"*, Section VIII, Pressures vessels, Division 2, Alternative rules, 225-230, 1983

[11] Kohnke, P., *"ANSYS users manual-Volume IV-Theory"*, Swanson Analysis Systems, Houston, 1992

ANALYSIS OF MULTIPLE CRACK PROBLEMS IN FINITE STRUCTURES

Michael R. Kratochwill
Tufts University
Department of Mechanical Engineering
Medford, MA 02155

Robert Greif
Tufts University

Christophe Mauge
Radionics, Inc.
22 Terry Avenue
Burlington, MA 01803

ABSTRACT: The problem of multiple crack interaction and the calculation of stress intensity factors is of great importance in failure analysis. Most solutions that are available are associated with infinite domains. Solutions for finite structural elements are more difficult to obtain and are often done by specialized techniques. The present technique is a general method for finite bodies which combines the finite element technique with the crack interaction method developed by Kachanov. The finite element method uses four noded quadrilateral isoparametric elements with stresses evaluated at nine Gaussian integration points. The Kachanov method for multiple crack problems assumes that the tractions on one crack is generated from the uniform average tractions of all other cracks. The combination of these two methods in an iterative technique produces the finite element alternating method for a finite plate. The effects of crack interaction and boundary distance are predicted for a variety of multiple crack problems. Additionally, a study is done of the effect of stiffeners on the multiple crack problem. In general it is shown that the stress intensity factors increase as the boundary of a plate is approached. However, the addition of a stiffener at the boundary is very effective at moderating the increase in stress intensity factors. The results of these studies should be of importance to designers and engineers concerned about failure prevention.

KEY WORDS: Crack interaction; Failure prevention; Finite Element Alternating Method (FEAM); Kachanov Method; Stress Intensity Factor (SIF).

INTRODUCTION: Design engineers often require analytical tools for evaluating structural damages that will influence the function of their design over time. Many factors contribute to the failure of structural components. Most notably the effects of crack interaction on the structural integrity of a design is vital to its overall success. The evaluation of the stress intensity factors (SIF) which directly influence crack propagation, is critical to the analysis of crack involvement on a structure. Many techniques exist for evaluating solutions for unbounded crack problems. However, practical structural design requires the evaluation of crack interaction within a finite area. Boundary interaction plays a critical role in the calculation of the SIF of cracks and consequently the integrity of the structure. While the finite element method (FEM) is an effective numerical technique for analyzing bounded crack problems, it is largely dependent upon the use of very fine and accurate element meshing. In order to optimize FEM, analytical solutions for infinite domain idealized structures are utilized. The finite element alternating method (FEAM) combines the advantages of FEM with the Kachanov method [1] for multiple cracks in *infinite* domains, to establish the solution for the problem of multiple cracks in a *finite domain.*

FINITE ELEMENT ALTERNATING METHOD: FEAM involves the iterative superposition of the finite element solution of a bounded uncracked plate with the analytical solution of an infinite two-dimensional plate with a crack subjected to external loadings. This method employs the Kachanov method for evaluating the analytical solutions for multiple arbitrarily oriented cracks in an infinite plate. FEAM iterates between this solution and the FEM numerical stress solution for a finite body under arbitrary loading to satisfy boundary conditions. The overall procedure involves the following steps:

(1) Model the finite plate geometry with the desired element meshing.
(2) Using FEM, solve this bounded plate with *no* cracks under the prescribed loading conditions. The residual stresses within each element are calculated at 9 Gaussian integration points.
(3) Using these stress values throughout the plate, calculate the tractions on the cracks of the finite plate.
(4) In order to satisfy crack face boundary conditions, the calculated normal and shear crack face tractions must be removed. To accomplish this, the opposite value of the crack face tractions are applied in the analytical solution for the infinite domain crack problem.
(5) The solution for the stress intensity factors are calculated by applying the crack tractions to the Kachanov method. This method evaluates SIF at each crack tip using the calculated crack face tractions. These crack tractions along each crack are averaged over the entire crack. This average crack traction is used to evaluate the effects of crack interaction. The initial crack tractions are added to the tractions resulting from crack interaction, to determine the overall stress values on each crack. These normal and shear stress values, $p(\xi)$ and $\tau(\xi)$, are used to calculate the SIF of each crack.
(6) Tractions are created on all boundary nodes of the finite plate elements. The tractions on each node are generated from the reversed crack face normal and shear tractions calculated in step 5, as well as the tractions resulting from crack interaction.
(7) Boundary tractions are converted to nodal forces on the boundary nodes of the finite element mesh.
(8) These boundary node forces with two degrees of freedom are applied to the uncracked finite body for the next iteration of the procedure.
(9) FEAM iterates between the FEM finite boundary solution and the infinite plate analytical solution until the tractions on the crack faces are below the prescribed tolerances. At this point the program sums the SIF calculated for each iteration to give the total crack tip SIF for the problem.

Finite Element Mesh: To evaluate the stresses that occur within a loaded finite plate the proper finite element mesh must be incorporated. For this work a 192 element plate mesh is used in conjunction with two different plate geometries. One plate has a width equal to 2 inches and height equal to 6 inches. The second plate has a width equal to 2 inches and height equal to 4 inches. The 192 element mesh involves 24 rows and 8 columns of elements (Figure 1). Each element within a meshed plate is of equal size.

Calculation of Element Stresses: For this work a four noded quadrilateral isoparametric element is used for evaluating element stresses. This element is characterized by bilinear shape functions where element displacements vary linearly. The shape functions of these elements are used to determine the element coordinates and displacements. The global element displacements are determined from the element shape functions and nodal displacements.

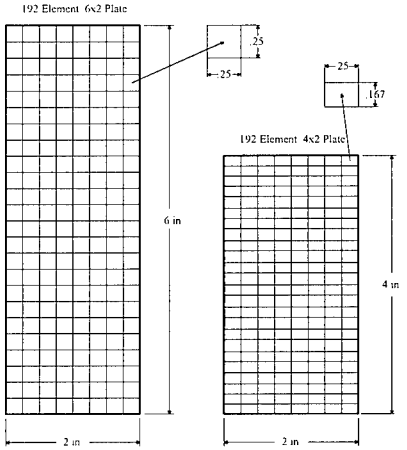


Figure 1: 192 Element Plate Meshes

The matrix of strain displacement [B] is used to relate global strains to local displacements. Coordinate transformation of derivatives is attained using the Jacobian operator [J] which contains the derivatives of the global coordinates with respect to the local coordinates (ξ,η). The stiffness matrix [K], which relates global nodal forces and displacements is derived from the strain energy of the body and has the form

$$[K] = t \cdot \iint [B]^T \cdot [C] \cdot [B] \cdot \det[J]\, d\xi\, d\eta \qquad (1)$$

where t is the element thickness and [C] is the matrix of elastic properties.

Using *Gaussian Quadrature* the elemental stresses are calculated at nine points within each element. These stresses within each element provide the overall stress distribution for the entire finite plate. The location and values of these stress points provide the data for solving the analytical equations of the infinite plate problem to determine the crack tip SIF.

Calculation of Stress Intensity Factors: The finite plate residual stresses calculated by FEM present tractions along the cracks of the infinite problem. From the plate stress data, stresses are evaluated at thirteen evaluation points along the crack (which divide the crack into twelve equal subdivisions) using a bivariate quadratic interpolation routine. From these crack point stresses (σ_x, σ_y, τ_{xy}) the equivalent normal and shear crack face tractions are calculated at each point using equations of standard theory of elasticity. The normal and shear stresses calculated at the thirteen crack evaluation points are used to evaluate the stress distribution along the entire crack. A third order degree polynomial is fit to the normal and shear crack stresses using a least squares method. The average normal and shear tractions along the crack are calculated from the stress distribution polynomial coefficients for each crack which are then used by the Kachanov method to analyze crack interaction and resultant crack SIF. To calculate the influence of the crack stresses upon the boundary nodes, Westergaard stress function equations are incorporated. The stress function equations are derived from the crack stress distribution polynomial coefficients. These functions are integrated to calculate the overall normal and shear boundary node stresses (σ_x, σ_y, τ_{xy}).

KACHANOV METHOD: Under this method the problem of a linear elastic solid with uniform stress σ^∞ containing N interacting cracks (unit normals n_i) is replaced by the equivalent problem: crack faces are loaded by $\boldsymbol{t} = -\boldsymbol{n}_i \cdot \sigma^\infty$ and stresses vanish at infinity [1,2]. This enables us to represent the problem as a superposition of N sub problems containing only one crack each but loaded by unknown tractions. The ith crack is loaded by

$$t_i(\xi) = t + \sum_k \Delta t_{ki}(\xi) \tag{2}$$

where Δt_{ki} is the traction induced by the isolated kth crack along the ith crack site and ξ runs $\pm\, a$ on any given crack. The key simplifying assumption of the Kachanov Method is that the traction $\Delta t_{ki}(\xi)$ is generated by the kth crack loaded by a uniform average traction $\langle t_k \rangle$ where the notation $\langle\ \rangle$ represents taking the average. The assumption neglects the impact on the ith crack of the traction non-uniformities $t_k - \langle t_k \rangle$ on the kth crack. As a result we can represent the normal and shear tractions on a typical crack defined as crack 1 as [1,3,4,5]

$$\begin{aligned} p_1(\xi) &= p_1^\infty + n_1 \cdot [\sigma_2^n(\xi)\langle p_2\rangle + \sigma_2^\tau(\xi)\langle \tau_2\rangle \cdots + \sigma_N^n(\xi)\langle p_N\rangle + \sigma_N^\tau(\xi)\langle \tau_N\rangle] \cdot n_1 \\ \tau_1(\xi) &= \tau_1^\infty + n_1 \cdot [\sigma_2^n(\xi)\langle p_2\rangle + \sigma_2^\tau(\xi)\langle \tau_2\rangle \cdots + \sigma_N^n(\xi)\langle p_N\rangle + \sigma_N^\tau(\xi)\langle \tau_N\rangle] \cdot (\mathrm{I} - n_1\, n_1) \end{aligned} \tag{3}$$

where p_1^∞ and τ_1^∞ are the normal and shear tractions induced on the first crack by the remote loading. The values σ_i^n and σ_i^τ are the stress fields generated on crack 1 by the ith crack with normal n and shear τ loading. If we take the average of Equation (3) over the length of crack 1 the following results are obtained,

$$\begin{aligned} \langle p_1\rangle &= p_1^\infty + \Lambda_{21}^{nn}\langle p_2\rangle + \Lambda_{21}^{n\tau}\langle \tau_2\rangle + \cdots \Lambda_{N1}^{nn}\langle p_N\rangle + \Lambda_{N1}^{n\tau}\langle \tau_N\rangle \\ \langle \tau_1\rangle &= \tau_1^\infty + \Lambda_{21}^{n\tau}\langle p_2\rangle + \Lambda_{21}^{\tau\tau}\langle \tau_2\rangle + \cdots \Lambda_{N1}^{n\tau}\langle p_N\rangle + \Lambda_{N1}^{\tau\tau}\langle \tau_N\rangle \end{aligned} \tag{4}$$

where Λ are termed the *transmission factors* and characterize the effect of the average tractions from one crack to another. As an example, $\Lambda_{21}^{n\tau}$, refers to the average *shear* (τ) on crack 1 caused by the *normal* (n) uniform load on crack 2 [1,3,4,5]. For all N cracks, Equations (4) form a system of 2N linear algebraic equations that can be solved to obtain the average *normal* $\langle p_k\rangle$ and *shear* $\langle \tau_k\rangle$ tractions by

$$(2\delta_{ik}\cdot \mathrm{I} - \Lambda_{ik}) \cdot \langle t_k\rangle = t_k^\infty \tag{5}$$

where $\langle t_k\rangle$ represents $\langle p_k\rangle$ and $\langle \tau_k\rangle$, and δ_{ik} is the Kronecker delta. If we substitute the average tractions $\langle p_i\rangle$ and $\langle \tau_i\rangle$ into Equations (3) the actual tractions $p_k(\xi)$ and $\tau_k(\xi)$ are found.

Using the actual tractions, the stress intensity factors at the crack tips are calculated from Equation (6). Where K_I & K_{II} are the respective mode I and mode II SIF and a is the half length of the crack.

$$\begin{Bmatrix} K_I \\ K_{II} \end{Bmatrix} = \frac{1}{\sqrt{\pi \cdot a}} \int_{-a}^{a} \begin{Bmatrix} p(\xi) \\ \tau(\xi) \end{Bmatrix} \cdot \sqrt{\frac{a+\xi}{a-\xi}}\, d\xi \qquad (6)$$

The Kachanov Method provides a very efficient tool for acquiring accurate analytical solutions to crack interaction problems. The combination of this crack interaction technique with FEM provides a powerful, accurate and efficient tool for evaluating complicated crack interaction problems in a finite plate.

RESULTS & DISCUSSION: All problems are considered plane stress with unit thickness, E = 30000.0 psi, υ = .3, and loaded by σ_1 = 1.0 psi. A FORTRAN based computer code was created to perform FEAM. This code is run on a DEC Alpha farm network which enables the code to run fast and efficiently with even the most complex problems taking no more than 5 seconds of cpu time to execute.

Single Center Crack: A simple finite boundary crack problem is run to demonstrate the validity and applicability of the program for performing FEAM. This problem consists of a single crack located at the center of a 6x2 finite plate subjected to uniform tensile stress at the upper and lower boundaries. The plate width ($2b$) is equal to 2 inches and the plate height ($2h$) equals 6 inches. This problem consists of pure symmetric mode I loading upon the crack resulting in equal SIF (K_I) at each tip of the crack. The SIF values (K_I) are calculated for crack lengths ($2a$) varying from 0.2 to 1.6 inches long. The are compared to results for identical problems from Isida[6] which are accurate to within 0.1%.

FEAM achieves exact results for 5 separate crack lengths ($2a$ = 0.2 to 1.0). For the crack lengths $2a$ = 1.2 and 1.4, the calculated SIF have errors less than .1%. For the longest crack, $2a$ = 1.6, the calculated SIF error is only .9% from the exact value. The SIF value for the longest crack is very accurate considering the relatively close proximity of the crack tip to the plate boundary. This distance, δ_B = 0.2, represents a region where a significant stress variation occurs per unit area. The stress deviations that occur within this small region can be effectively evaluated by FEAM because stresses are evaluated at 9 points within each element. In all, the 192 element mesh contains a total of 1728 stress points throughout the finite plate. Consequently, FEAM rapidly converges towards the final solution in less than 10 iterations.

Parametric Study of Crack Growing Towards Boundary: To demonstrate the influence of the plate boundary upon a crack tip, various problems are performed where the crack tip grows towards the boundary. These problems consist of a single crack with its center offset from the vertical centerline of the plate (Figure 2). All problems involve a 6x2 plate loaded under tension at the upper and lower boundaries. The problem deals with the crack center offset a distance of e = 0.5 inches from the vertical centerline of the plate, resulting in the parameter e/b = 0.5. The crack lengths ($2a$) are incrementally increased resulting in a decrease of the distance between the crack tip and plate boundary (δ_B). *Normalized* SIF for mode I loading are obtained for the outer tip (tip B), as well as the inner tip (tip A). Normalized values are given as

$$F_I = K_I \Big/ \sigma\sqrt{\pi a} \qquad (7)$$

The results for F_{IA} and F_{IB} are detailed in Table I and are compared to graph values for the same problem performed by Isida [6]. It is important to note that these interpolated reference results are accurate to 2% and are not considered to be exact. Any deviation from these results are considered as differences. It can be seen that as the distance between the crack tip and plate boundary decreases the calculated results for SIF at the outer crack tip become less accurate. For very close distances between the crack tip and plate boundary, δ_B< 0.2, the largest difference from the reference values is only 2.3% (δ_B = 0.1). With the validity of the program established, extensive evaluation of complex crack interaction problems can be attempted. The *Kachanov method* is employed for multiple crack problems in order to capture the effects of crack interaction upon the calculation SIF for each crack within the finite plate.

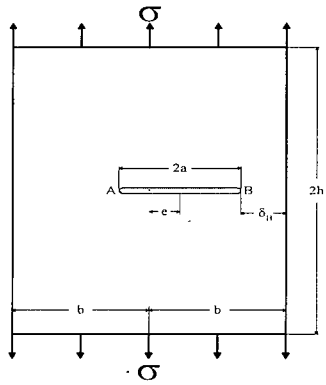


Figure 2: Crack Growing To Boundary

Table I: Single Crack Growing To Boundary,
(e/b = 0.5, $2h$ = 6 inches, $2b$ = 2 inches)

Crack Lgth. $2a$ (in.)	δ_B	FEAM SIF F_{IA}	Isida SIF F_{IA}	% Dif	FEAM SIF F_{IB}	Isida SIF F_{IB}	% Dif
.20	.40	1.01	1.01	0	1.02	1.02	0
.30	.35	1.04	1.04	0	1.05	1.05	0
.40	.30	1.06	1.07	.9	1.09	1.09	0
.50	.25	1.09	1.10	.9	1.16	1.17	.9
.60	.20	1.13	1.15	1.7	1.25	1.27	1.6
.70	.15	1.26	1.23	2.4	1.41	1.44	2.1
.80	.10	1.37	1.33	3.0	1.69	1.73	2.3

Two Collinear Cracks: This crack interaction problem consists of two collinear cracks symmetric about the y-axis of a 6x2 plate (Figure 3). This plate is loaded under tension at the boundaries. *All cracks have length $2a$ = 0.4 inches* and are aligned along the centerline of the plate. Problems are analyzed for distances between crack centers (δ_c) ranging from 0.6 to 1.4 inches. Normalized mode I SIF values (F_I) are calculated at both the inner (A) and outer (B) tips of each crack and compared to results for two collinear cracks within a strip plate having a finite width and infinite height. Since we are investigating a symmetric problem, the SIF values are equal for the corresponding tips of each crack. The results are presented in Table II as F_{IA} and F_{IB} and are compared to results calculated by Gupta and Erdogan [7].

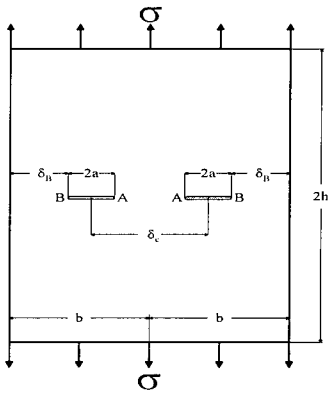


Figure 3: Two Collinear Cracks

Table II: Two Collinear Cracks with
Varying Center Distances
($2a$ = 0.4 in., $2h$ = 6 in., $2b$ = 2 in.)

Crack Lgth $2a$ (in.)	δ_c	FEAM SIF F_{IA}	G&E SIF F_{IA}	% Dif	FEAM SIF F_{IB}	G&E SIF F_{IB}	% Dif
0.4	0.6	1.167	1.17	.4	1.106	1.11	.9
0.4	0.8	1.108	1.11	.2	1.088	1.09	.7
0.4	1.2	1.108	1.09	.9	1.103	1.12	1.9
0.4	1.4	1.159	1.12	2.6	1.163	1.22	5.3

The results for the distances between crack centers δ_c = 0.6, 0.8 and 1.2 are very accurate with respect to the reference values. The problem with an overall crack center offset of δ_c = 1.4 achieves results which differ 2.6% for the inner crack tips (A) and 5.3% for the outer crack tips (B). The result for the inner tips is quite good considering the complexity of problem and the addition of crack interaction to the problem domain. The outer tip results are considerably more difficult for this problem, *for both FEAM and the analytical results of Gupta & Erdogan,* since the distance between the crack tip and boundary, δ_B , is only .1 inches. The large stress deviation in this region complicates the calculation of SIF for the outer tips.

Two Inclined Cracks: The next problem is designed to evaluate both mode I and mode II loading applied to multiple cracks. This problem consist of two inclined cracks, symmetric about the y-axis, within a 4x2 plate under uniform tension at the boundaries (Figure 4). The distance between the two crack centers is δ_c = 0.55 inches, with each crack length ($2a$) equal to 0.5 inches. The angular orientation of the cracks are as follows: crack 1 rotated at an angle θ = +45° and crack 2 rotated at an angle θ = -45°. Mode I and mode II SIF values (F_I ,F_{II}) are calculated at each tip for both cracks. These values are shown in Table III and referenced to the identical problem performed by Chen and Chang [8].

The SIF values are extremely accurate for F_I (mode I) at the outer (B_1 & B_2) and inner tips (A_2 & A_2) having a difference of only .2 to .3%. The results are also quite good for F_{II} (mode II) at both the outer tips and inner tips, with the largest difference from the reference results being 1.6%. For this problem, crack tips A_1 and A_2 are in close proximity to one another which creates extreme instances of stress interaction between the two cracks which in turn plays a tremendous role in calculating the crack tractions induced on one crack by the other. The ability to capture the complex events which occur in such a small area is critical to the success of FEAM.

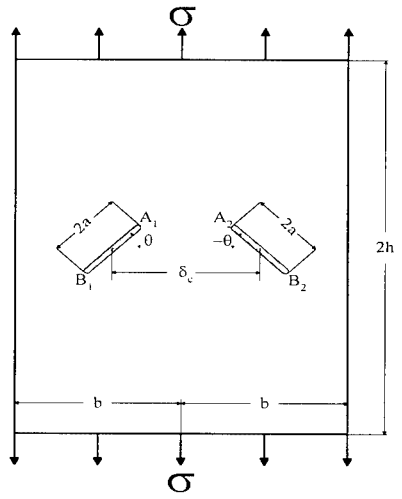


Figure 4: Two Inclined Cracks

Table III: Two Inclined Cracks, $\theta = \pm 45°$
($2a$ = 0.5, $2h$ = 4, $2b$ = 2, δ_c = 0.55 inches)

MODE I

Tip	FEAM SIF F_I	Chen SIF F_I	% Diff.
B_1	.519	.518	.2
A_1	.593	.595	.3
A_2	.593	.595	.3
B_2	.519	.518	.2

MODE II

Tip	FEAM SIF F_{II}	Chen SIF F_{II}	% Diff.
B_1	.543	.552	1.6
A_1	.506	.511	.9
A_2	-.506	-.511	.9
B_2	-.543	-.552	1.6

Two Parallel Offset Cracks: This problem consists of two parallel cracks of equal length ($2a$ = 0.2) that are *not* aligned in the tensile stress direction. These cracks lie within a 6x2 plate loaded under tension at the boundaries (Figure 5). Both cracks are oriented perpendicular to the tensile stress direction with the crack centers offset from each other a distance δ_c. The vertical displacement between each crack, f, as well as the crack center offset, δ_c, are varied to change the actual distance between crack centers (d) resulting in various interaction parameter values ($2a/d$). Normalized SIF values are calculated at the inner tips of each crack (A tips). Due to the symmetric nature of the crack orientations, the SIF at both of these tips are equal and are presented as one value (F_{IA}). The results are presented in Table IV and compared to values for problems analyzed within an *infinite* plate by Isida using a Laurent series expansion with accuracy of 2% [9].

Statistically, all the results for this problem are within the prescribed 2% tolerance. Overall, these results reconfirm the power and flexibility of FEAM for analyzing complex finite boundary crack interaction problems. With the validity and applicability of FEAM established with respect to comparative problems, new finite boundary crack interaction problems can be pursued. FEAM has the ability to evaluate many problems within a finite boundary that until now have only been performed using an *infinite* plate domain.

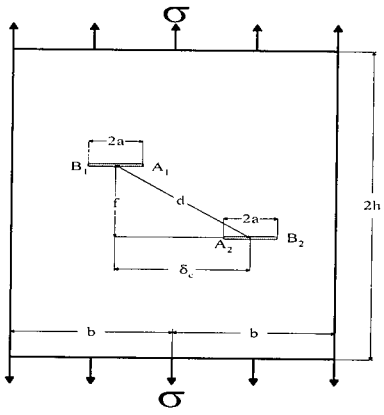


Figure 5: Two Parallel Cracks with Offset

Table IV: Two Parallel Offset Cracks, $\delta_c / f = 2.0$
($2a = 0.2$, $2h = 6$, $2b = 2$)

δ_c	f	$2a/d$	FEAM SIF F_{IA}	∞ Plate SIF F_{IA}	% Dif f
.700	.350	.25	1.027	1.010	1.7
.500	.250	.35	1.042	1.025	1.7
.400	.200	.45	1.061	1.050	1.0
.300	.150	.60	1.104	1.090	1.3
.260	.130	.69	1.130	1.120	.9
.240	.120	.75	1.141	1.130	1.0
.220	.110	.81	1.143	1.135	.7
.200	.100	.90	1.122	1.120	.2

Two Parallel Offset Cracks Growing Towards Boundary: Another practical engineering problem to analyze is the situation where offset cracks grow towards the plate boundary. This problem consists of two parallel cracks that are placed within a 6x2 plate loaded under tension at the boundaries (Figure 6). The crack centers are horizontally offset a distance $\delta_c = 0.2$ inches and are separated vertically a distance of $d = 0.25$ inches. The crack lengths ($2a$) are incrementally increased in order to decrease the distance between the outer crack tips (B) and the boundary. By gradually increasing the lengths of each crack, the problem is designed to simulate crack growth. Consequently, the effects of crack propagation upon the SIF of the tips is investigated. Normalized mode I SIF values are calculated at both the inner and outer tips (F_{IA} and F_{IB}) for various outer tip to boundary distances (δ_B) and presented in Table V.

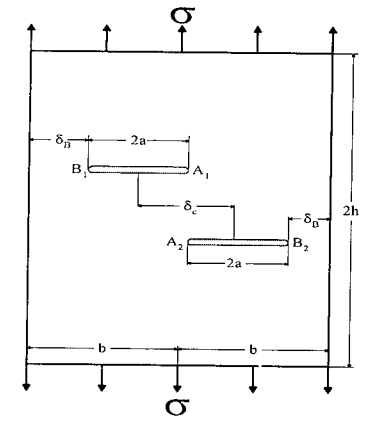


Figure 6: Two Parallel Offset Cracks Growing Towards Boundary

Table V: Two Parallel Offset Cracks Growing Towards Boundary
($\delta_c = 0.2$, $d = 0.25$, $2h = 6$, $2b = 2$)

Crack Length $2a$	Tip B to Boundary δ_B	Tip A SIF F_{IA}	Tip B SIF F_{IB}
.20	.80	.968	1.038
.40	.70	.754	1.091
.60	.60	.681	1.145
.80	.50	.739	1.246
1.00	.40	.888	1.434
1.20	.30	1.163	1.787
1.40	.20	1.742	2.556

As the distance between the outer crack tips and boundary (δ_B) decrease, the SIF at the outer tip gradually increases. The results demonstrate that the increase in magnitude of F_{IB} is proportional to the decrease of δ_B, down to a distance of 0.3 inches. Once $\delta_B \leq 0.3$ the SIF value increases significantly. As the cracks grow from 0.2 to 0.6 inches, the inner tip SIF (F_{IA}) *decrease*. When the cracks initially grow, the inner

portions of each crack begin to overlap and shield each other from the opposing tensile. Once the cracks grow past 0.6 inches, the inner tips begin to feel the effects of the plate boundary which has a greater influence upon the crack tip stresses than the effects of shielding. Consequently, the SIF at the inner tips (F_{IA}) significantly *increase* once they reach this region.

Single Center Crack with Plate Stiffeners: As previously discussed, the need to investigate plate stiffeners is vital to any stress analysis technique. This next problem consists of a single center crack within a 6x2 plate loaded under tension at the boundaries (Figure 7). Plate stiffeners are fastened to the left and right sides of the plate which restrict the *y* degree of freedom nodal displacements that can occur on these sides. Normalized mode I SIF values are calculated for various crack lengths ($2a$). For all crack lengths evaluated, the SIF values (F_I) are less than 1% different from results for an identical problem performed by Cartwright [10]. Table VI compares the results for this problem to the SIF values (normalized form) calculated for the identical single center crack problem *without* plate stiffeners previously presented.

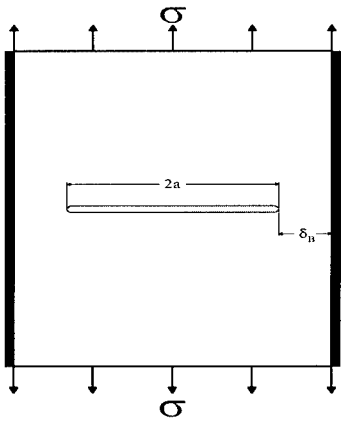


Figure 7: Single Center Crack with Stiffeners

Table VI: Single Center Crack with Stiffeners
($2h = 6$, $2b = 2$)

Lgth $2a$	SIF Stiffened Plate F_I	SIF No Stiffeners F_I	% Decrease Due to Stiffeners
1.6	.086	1.801	95.2
1.4	.095	1.488	93.6
1.2	.103	1.302	92.1
1.0	.108	1.187	90.9
.8	.114	1.110	89.7
.6	.118	1.058	88.8
.4	.121	1.024	88.2
.2	.122	1.005	87.8

Comparison of the SIF results for the crack within a *stiffened* plate to the results of the same crack within an *unstiffened* plate demonstrate the effectiveness of the plate stiffeners. The results show that the addition of stiffeners to the plate reduce the crack SIF by at least 87.8%. For all problems without plate stiffeners the SIF drastically increases with the lengthening of a crack due to the decrease between the crack tip and boundary. However, with the addition of plate stiffeners the overall stresses surrounding a crack tip are redirected to reduce the SIF at that tip which limits crack propagation.

Two Parallel Offset Cracks Growing Towards Stiffened Boundary: The last stiffened plate problem evaluated consists of two parallel cracks growing towards a stiffened plate boundary. The cracks are placed within a 6x2 plate with stiffeners placed along the sides of the plate. This plate is loaded under tension at the (Figure 8). The crack centers are horizontally offset a distance $\delta_c = 0.2$ inches and are separated a vertical distance $d = 0.25$ inches. In order to simulate crack growth, the crack lengths ($2a$) are incrementally increased which decreases the distance between the outer crack tips (B) and the stiffened side boundaries. Normalized mode I SIF values are calculated at both tips (F_{IA} and F_{IB}) for various distances between the outer crack tips and the stiffened boundary (δ_B). These values are presented in Table VII and are compared to the identical problems *without* stiffeners analyzed earlier. The addition of stiffeners to the plate boundary significantly reduce the SIF at the crack tips. Figure 9 demonstrates the effectiveness of the stiffeners for limiting the stress intensity factors for both the inner and outer tips. When the problem consists of *no* stiffeners, the SIF (F_{IA} and F_{IB}) rise at a rapid rate as the cracks grow and the tips move closer to the boundary. When the stiffeners are added to the boundary, the corresponding SIF are dramatically reduced.

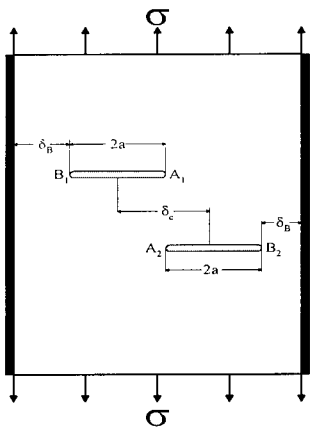


Figure 8: Two Parallel Offset Cracks Growing Towards Stiffened Boundary

Table VII: Two Parallel Offset Cracks Growing Towards Stiffened Boundary ($2h = 6$, $2b = 2$, $\delta_c = 0.2$, $d = 0.25$)

$2a$	δ_B	Stiff. Plate F_{IA}	No Stiff. F_{IA}	% Dec	Stiff. Plate F_{IB}	No Stiff. F_{IB}	% Dec
.20	.80	.956	.968	1.2	1.026	1.038	1.2
.40	.70	.706	.754	6.3	1.042	1.091	4.5
.60	.60	.569	.681	16.4	1.026	1.145	10.4
.80	.50	.526	.739	28.8	1.001	1.246	20.0
1.00	.40	.514	.888	42.1	.998	1.434	30.4
1.20	.30	.512	1.163	55.9	.990	1.787	44.6
1.40	.20	.512	1.742	70.6	.984	2.556	61.5

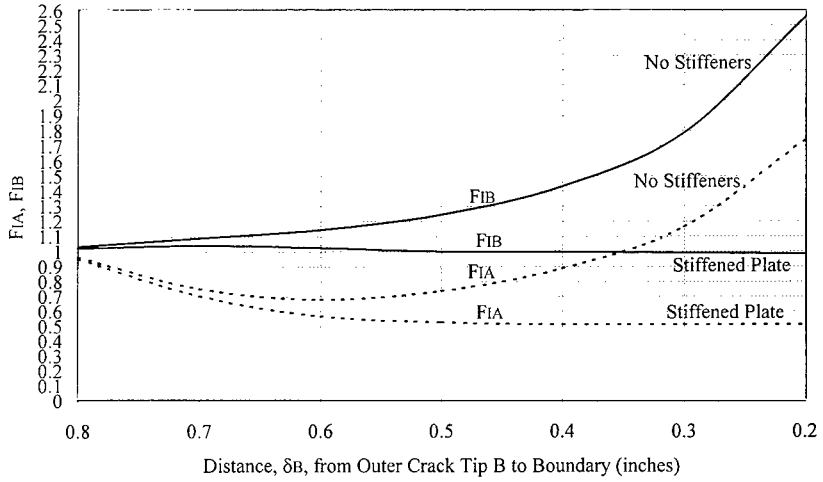


Figure 9: F_I vs. δ_B for Two Parallel Offset Cracks Moving Towards Boundary Comparison of Stiffened Boundary to Non-Stiffened Boundary

The *stiffened* plate reacts as if it is semi-infinite, having no lateral boundary interaction. Consequently, as the crack tips move closer to the *stiffened* boundary the SIF do not change as dramatically as when there are *no* stiffeners. These stiffeners constrain the displacement of the finite element nodes. The calculation of residual stresses within a tensile loaded plate are dependent upon nodal displacements. When the nodal displacements are constrained, the resulting stresses within the plate are significantly reduced. Accordingly, smaller stresses within the plate result in smaller stresses along the cracks. Therefore, the SIF of the cracks are reduced. Overall, the data proves that the addition of stiffeners along the sides of a finite plate can be beneficial for limiting stresses around the crack tip. These stiffeners are effective for diverse crack

interaction problems. As illustrated by the problems presented, plate stiffeners have many practical applications which aid design engineers in the prevention of crack growth. If placed properly along a structural member, a stiffener can limit crack propagation by reducing displacements and forces in the region surrounding the crack.

CONCLUSION: Design engineers often require analytical tools for evaluating structural damages that influence the function of their design. Over time multiple cracks can contribute to the overall failure of a structural component. In order to investigate the effects of crack interaction upon a structural member the evaluation of the crack stress intensity factors (SIF) is paramount. These crack stress intensity factors play a large role as fracture correlation parameters in current engineering practice. Mechanical structures often display fatigue cracks which must be closely monitored and evaluated in order to prevent failures which could compromise the performance of that structure. Many structural components involve stiffening plates, different screw and welded fasteners, and complex geometrically distributed features all within a finite boundary. Hence, a numerical technique which can evaluate fracture related parameters within a finite area is quite valuable to many design engineering disciplines. The establishment of the finite element alternating method (FEAM) as such a solution technique is the objective of this work. The Kachanov method is incorporated to effectively analyze the involvement of crack interaction upon a loaded structure in order to optimize the FEAM computer program. This computer program efficiently performs FEAM to calculate accurate stress intensity factors for diverse finite boundary crack interaction problems.

The investigation of various problems present interesting conclusions concerning the analysis and prevention of crack propagation. First, as the distance between the plate boundary and crack tip becomes smaller the resulting crack stress intensity factor increases. The boundary dramatically influences the stresses that occur along a crack when the crack tip approaches the boundary. Similarly, as crack tips move closer together the influence of crack interaction upon the entire problem domain significantly magnifies. Application of stiffeners along the boundary of a plate prove to be practical tools for limiting crack stress intensity factors. These stiffeners reduce the displacements near the boundary which in turn decreases the strong influence of the boundary. These stiffeners drastically reduce the stress intensity factors when applied to the region proximal to a crack. Consequently, stiffeners often are applied in order to decrease the propagation of fatigue cracks in structural components.

Bibliography

[1] M. Kachanov. Elastic solids with many cracks: A simple method of analysis. *International Journal of Solids and Structures,* 23:23-44,1987.

[2] C. Mauge and M. Kachanov. Interacting arbitrarily oriented cracks in anisotropic matrix: Stress intensity factors and effective moduli. *International Journal of Fracture,* 58:69-74, 1992.

[3] T.S. Moore. *Analysis of Two-Dimensional Interacting Crack Problems Using a Finite Element Alternating Method.* Master's thesis, Tufts University, 1993.

[4] C.P. Mauge. *Analysis of Linear Elastic Anisotropic Solids With Randomly Oriented Cracks.* Ph.D. thesis, Tufts University, 1993.

[5] M.R. Kratochwill. *Evaluation of Stress Intensity Factors for Crack Interaction Problems Using a Finite Element Alternating Method.* Master's thesis, Tufts University, 1996.

[6] H. Tada, P. Paris, and G. Irwin. *The Stress Analysis of Cracks Handbook.* Del Research Corp., 1985.

[7] G.D. Gupta and F.Erdogan. The problem of edge cracks in an infinite strip. *Journal of Applied Mechanics,*1001-1006, 1974.

[8] W. Chen and C. Chang. Analysis of two-dimensional mixed mode crack problems by finite element alternating method. *Computers & Structures,* 33:1451-1458, 1989.

[9] M. Isida. Analysis of Stress Intensity Factors for Plates Containing Randomly Distributed Cracks. *Transactions of Japanese Society of Mechanical Engineers,* 35: 1815-1822, 1969.

[10] D. Cartwright. Stiffened Sheets. In Y. Murakami, editor-in-chief, *Stress Intensity Factors Handbook,* 2.1.1. Pergamon, 1987.

SPACE ENRICHMENT METHOD FOR FINITE ELEMENT APPROXIMATION OF MIXED BOUNDARY PROBLEM IN TWO-DIMENSIONAL ELASTICITY

Yehuda Volpert[1]
Washington University in St. Louis
Center for Computational Mechanics
Campus Box 1129, One Brookings Drive
St. Louis, MO 63130

Barna A. Szabó
Washington University in St. Louis
Center for Computational Mechanics
Campus Box 1129, One Brookings Drive
St. Louis, MO 63130

Abstract: This paper is concerned with numerical approximation of the solution of elliptic boundary value problems with Winkler type boundary conditions in two-dimensional elasticity. An extension of the p-version of the finite element method, called the space enrichment method, is described and illustrated by examples. The nature of the singularities is discussed and special basis functions, capable of capturing the singular behavior, are introduced. The accuracy of the space enrichment method is evaluated against numerical solution that of the hp-extension. The two methods exhibit very close agreement.

Key Words: Finite element; p-version; elasticity; mechanical contact; singular functions; numerical solutions.

INTRODUCTION: This research is motivated by the necessity to construct accurate approximations for the numerical solution of mechanical contact problems. In particular, the displacement and stress fields in contact region are of interest. The numerical formulation of the contact problem was initiated by Signorini [1]. The variational theory of mechanical contact was developed by Duvaut and Lions [2] who provided rigorous proofs of existence and uniqueness of the solution.

The two variational formulations most frequently used in connection with mechanical contact are the penalty method and the augmented Lagrangian method [3,4]. In both of these methods the solution depends on the penalty coefficient. The physical meaning of this coefficient can be interpreted as a spring stiffness coefficient of the Winkler type boundary. The Winkler foundation is essentially a distributed linear spring which correlates normal tractions with relative normal displacements and is often used for the description of the boundary conditions in the contact region. The problem with spring boundary condition is also known as mixed or Neumann-Robin boundary value problem. This approach is also used in numerical approximations of contact. However, an adequate

[1] Currently: AlliedSignal 111 S. 34th Street Mail Stop 301-124 Phoenix, AZ 85072

treatment of spring boundary conditions has not been addressed in the past. Only recently Costabel and Dauge [5] established the nature of the solution for a scalar elliptic boundary value problem in $\mathbf{R}^2$.

Accurate and efficient numerical methods for the approximation of the solutions of the vector elliptic boundary value problems in $\mathbf{R}^2$ are of great interest because the solution of contact problems is very important. The solution is smooth everywhere in the domain except in the points where the spring terminates and thus an abrupt change in the boundary conditions is introduced. In these points the solution exhibits *nearly singular* behavior. By classification of the solutions given in [6], the mixed boundary value problem belongs in category $\mathbf{B}$. One way to approach solutions which belong in category $\mathbf{B}$ is to use hp-refinement. Meshes graded in geometric progression with two or more layers of geometrically graded elements with a common factor of 0.15 towards the singular point provide very accurate numerical approximations of the solution. The only disadvantage of the hp-method is that solution dependent mesh refinement is involved which carries a large overhead in non-linear solutions. In this work another approach, called the space enrichment method is investigated. The main idea of the space enrichment method is to augment the polynomial finite element space with functions capable of closely representing the exact solution in the neighborhood of the singular points. The nature of the solution and the type of functions which have to be incorporated in the finite element space are discussed next.

THE NATURE OF THE SOLUTION: Consider a semi-infinite elastic domain, part of which is subject to a spring boundary of the Winkler type on boundary segment Γ_s:

$$T_n - k_s(\delta_n - u_n) = 0 \tag{1}$$

where T_n is the normal stress, u_n is the displacement component normal to the boundary, k_s is the stiffness of the spring and δ_n is an imposed displacement. In the case of stiff spring (k_s large) the solution is close to the case of imposed displacement. This is clearly visible from the potential energy expression:

$$\Pi = W + \frac{1}{2}\int_{\Gamma_s} k_s(\delta_n - u_n)^2 ds - P \tag{2}$$

where Π is the potential energy, W is the strain energy, the integral expression is the strain energy of the spring and P is the potential of the external forces. For large k_s the integral expression must be close to zero, hence $\delta_n \approx u_n$. When k_s is small then the total potential energy of the system is given by

$$\Pi = \underbrace{W + \frac{1}{2}\int_{\Gamma_s} k_s u_n^2 ds}_{\text{First term}} + \underbrace{\frac{1}{2}\int_{\Gamma_s} k_s \delta_n^2 ds}_{\text{Second term}} - \underbrace{\left(P + \int_{\Gamma_s} k_s \delta_n u_n ds\right)}_{\text{Third term}}. \tag{3}$$

The first term in (3) represents the strain energy, including the effects of spring. Since k_s is small, the strain energy is close to the strain energy of the body without the spring. The second term is a constant which does affect the minimum of the potential energy. The third term is the potential of the external loads, including the effects of the spring. In this case $k_s\delta_n$ is a traction term, hence the effect of the spring boundary condition is similar to

the effect of an applied traction the magnitude of which is $k_s\delta_n$. Thus there are two limiting cases. In the case of stiff springs we will be interested in the homogeneous solution, i.e. $\delta_n = 0$ and hence $u_n = 0$. In the case of soft springs we will be interested in the particular solution, corresponding to an abrupt change in tractions.

Two extreme cases are considered: the first case is the case of a very stiff spring, and, the second case is the case of a very soft spring. The problem of two-dimensional elasticity is formulated in terms of Airy stress function, U. (For comprehensive explanation of this theory the reader is referred to Muskhelshvili [7].) The Airy stress function is a biharmonic function, which is given in terms of a complex variable $z = x + iy$ by

$$U = \Re\{\bar{z}\varphi(z) + \chi(z)\} \tag{4}$$

where $\varphi(z)$ and $\chi(z)$ are holomorphic functions, and $\overline{(\cdot)}$ denotes the conjugate variable or function. The displacement components are obtained from the formula

$$2\mu(u_x + iu_y) = \kappa\varphi(z) - z\overline{\varphi'(z)} - \overline{\psi(z)} \tag{5}$$

where $\psi(z) = \chi'(z)$. In this formula μ is the shear modulus, $\kappa = 3 - 4\nu$ for plane strain and ν is the Poisson's ratio. The complex representation of the components of the stress tensor is given by

$$\sigma_{xx} + \sigma_{yy} = 2[\Phi(z) + \overline{\Phi(z)}] \tag{6}$$

$$\sigma_{yy} - \sigma_{xx} + i\tau_{xy} = 2[\bar{z}\Phi'(z) + \Psi(z)] \tag{7}$$

where $\varphi(z) = \int \Phi(z)dz$ and $\psi(z) = \int \Psi(z)dz$.

The homogeneous part of the solution for the first limiting case, the case of a very stiff spring, $k_s \to \infty$, corresponds to a well known problem of a crack located along a line of symmetry in an infinite domain. The Airy stress function for this problem, given in [6], is

$$U = \Re\left(\bar{z}z^{\alpha} - \frac{(\alpha-1)\sin(\alpha-1)\pi}{(\alpha+1)\sin(\alpha+1)\pi} z^{\alpha+1}\right) \tag{8}$$

where $\alpha = \pm\frac{1}{2}, \pm\frac{3}{2}, \pm 2, \cdots$. Only those functions which have finite strain energy are considered, hence $\alpha > 0$. Consideration of the boundary value problem with a free-free boundary conditions at $\theta = 0$, and, fixed in the normal and free in the tangential direction for $\theta = \pi$, corresponds to the eigenfunctions which for any α are given by

$$u_x = \frac{A_\alpha r^\alpha}{2\mu}\{[\kappa + (\alpha-1)]\sin\alpha\theta - \alpha\sin(\alpha-2)\theta\} \tag{9}$$

$$u_y = \frac{A_\alpha r^\alpha}{2\mu}\{[\kappa - (\alpha-1)]\cos\alpha\theta + \alpha\cos(\alpha-2)\theta\}. \tag{10}$$

The coefficients A_α represent the amplitudes of the eigenfunctions. These eigenfunctions represent the natural displacement field, their derivatives represent the natural straining modes at the singular point.

Remark 1. *The solution of the problem in two-dimensional elasticity with Winkler type boundary conditions can be expressed in terms of analytic functions in vicinity of the singular point if one assumes that the spring coefficient is inversely proportional to the distance from the singular point:*

$$k_s = \frac{c}{r} \tag{11}$$

where c *is an arbitrary constant. In this case the eigenvalue problem resolves into transcendental equation:*

$$\alpha \tan(\pi\alpha) = -\frac{c}{4\mu}(\kappa+1) \tag{12}$$

from which it can be seen that the first eigenvalues bounded by

$$\tfrac{1}{2} < \alpha < 1 \tag{13}$$

where the lower bound corresponds to $c = \infty$, *the upper bound to* $c = 0$.

To obtain an approximation of the solution for the second limiting case, a classical problem of semi-infinite domain subjected to a uniform pressure distribution of intensity p over an interval $-a \le x \le a$, where x is the coordinate along the half-plane boundary, is considered. This problem is the first fundamental problem for the half-plane and the holomorphic functions for this case are given by Muskhelishvili:

$$\Phi(z) = \frac{p}{2\pi i}\ln\frac{z-a}{z+a} \tag{14}$$

and

$$\Psi(z) = -\frac{paz}{\pi i\left(z^2-a^2\right)}. \tag{15}$$

The characteristic part of the exact solution for the displacement components is then given by [8]:

$$2\mu u_x = (\kappa-1)\frac{pa}{\pi}\theta_2 + (1-\kappa)\frac{pr_1}{2\pi}(\theta_1-\theta_2)\cos\theta_1 - (1+\kappa)\frac{pr_1}{2\pi}\ln\frac{r_1}{r_2}\sin\theta_1 \tag{16}$$

$$2\mu u_y = (1+\kappa)\frac{pa}{\pi}\ln r_2 + (\kappa-1)\frac{pr_1}{2\pi}(\theta_1-\theta_2)\sin\theta_1 + (1+\kappa)\frac{pr_1}{2\pi}\ln\frac{r_1}{r_2}\cos\theta_1 \tag{17}$$

where the pairs $(r_1;\theta_1)$ and $(r_2;\theta_2)$ denote local polar coordinates with the origins on the on the half-plane boundary at $(a;0)$ and $(-a;0)$ respectively; the positive sense of the corresponding angles is chosen in clockwise direction and the angles are measured from the $x-axis$. It can be seen that the displacements are continuous, including the points $(\pm a;0)$.

Remark 2. *Let* $r_2 >> r_1$, *then the displacement components in the neighborhood of the point* $(a;0)$ *can be approximated by functions*

$$u_x = c_{x0} + c_{x1}r\theta\cos\theta - c_{x2}r\ln r\sin\theta \tag{18}$$

$$u_y = c_{y0} + c_{y1}r\theta\sin\theta + c_{y2}r\ln r\cos\theta \tag{19}$$

where r and θ *are the polar coordinates with the origin in this point and* c_{xk} *and* c_{yk}, $k = 0,1,2$ *are unknown coefficients which have to be determined.*

The nature of the solution in the neighborhood of the singular points was described and analytical results of the two limiting cases of the Winkler-type boundary conditions were presented. The solution of the mixed boundary value problem is likely to be somewhere between these two extremes. This suggests that, to be able to approximate the solution in vicinity of the '*nearly singular*' points, the span of the finite element space should be enlarged to include functions of the form $r\theta f(\theta)$, $r\ln rf(\theta)$ and $r^{\lambda_i} f(k_i\theta)$ where $f_i(\theta) = \sin(k_i\theta)$ or $\cos(k_i\theta)$, and, λ_i and k_i are fractional powers and coefficients. The proper selection of these values is very important for good approximation and is discussed in the section on numerical examples. The augmentation of the standard polynomial space with these functions is called the space enrichment method. The description of the space enrichment method is given next.

FINITE ELEMENT FORMULATION: To obtain the approximate solution to a problem in two-dimensional elasticity, the variational equations are derived by minimizing the potential energy, $\Pi(\vec{u})$, a quadratic functional, which by definition is

$$\Pi(\vec{u}) \equiv \frac{1}{2} B(\vec{u},\vec{u}) - F(\vec{u}). \tag{20}$$

The variational form of the governing equations is then given by:

$$B(\vec{u},\vec{w}) = F(\vec{w}) \quad \forall \vec{w} \in \overset{0}{E}(\Omega) \tag{21}$$

Where $B(\vec{u},\vec{w})$ is a *bilinear form* and $F(\vec{w})$ is a *linear functional* given respectively by

$$B(\vec{u},\vec{w}) = \iint_{\Omega} \left(\sigma_{xx}^{(u)}\varepsilon_{xx}^{(w)} + \sigma_{yy}^{(u)}\varepsilon_{yy}^{(w)} + \tau_{xy}^{(u)}\gamma_{xy}^{(w)}\right) t_z dxdy \tag{22}$$

$$F(\vec{w}) = \int_{\Gamma} \left(T_n w_n + T_t w_t\right) t_z ds \tag{23}$$

with no body forces present.

We seek a solution for a mixed boundary value problem for a circumstance in which the discontinuous data can be located at any point along the element edge. For this reason the solution vector is approximated by [8]

$$u_x = \sum_{i=1}^{n} N_i a_i + \sum_{i=1}^{m} \psi_i a_{2n+i} \tag{24}$$

$$u_y = \sum_{i=1}^{n} N_i a_{n+i} + \sum_{i=1}^{m} \psi_i a_{2n+m+i} \tag{25}$$

where N_i are the hierarchic shape functions defined on the standard quadrilateral element $\Omega_e^{(q)} = \{\xi,\eta \,|\, -1 \le \xi \le 1, -1 \le \eta \le 1\}$, and the ψ_i are the special basis functions defined in the polar coordinates with an origin in the *nearly singular* point. This formulation leads to a linear system of equations

$$\begin{bmatrix} K_{NN} & K_{N\psi} \\ K_{N\psi} & K_{\psi\psi} \end{bmatrix} \begin{Bmatrix} a_N \\ a_\psi \end{Bmatrix} = \begin{Bmatrix} r_N \\ r_\psi \end{Bmatrix} \tag{26}$$

where K_{NN} is the submatrix given in terms in terms of N-type basis functions, $K_{\psi\psi}$ represents the submatrix corresponding to the space enrichment functions, and, $K_{N\psi}$ represents the submatrix due to the coupling between the N-type and ψ-type basis functions. Each submatrix is composed of 4 other submatrices associated with the appropriate coefficients of the displacement components. For instance,

$$\left[K_{N\psi}\right] = \begin{bmatrix} K_{N\psi}^{11} & K_{N\psi}^{12} \\ K_{N\psi}^{21} & K_{N\psi}^{22} \end{bmatrix} \tag{27}$$

where, for example,

$$\left[K_{N\psi}^{12}\right] = \iint_{\Omega_e} \left(E_2 \frac{\partial N_i}{\partial x}\frac{\partial \psi_j}{\partial y} + E_4 \frac{\partial N_i}{\partial x}\frac{\partial \psi_j}{\partial x} + E_5 \frac{\partial N_i}{\partial y}\frac{\partial \psi_j}{\partial y} + E_6 \frac{\partial N_i}{\partial y}\frac{\partial \psi_j}{\partial x} \right) t_z dxdy \tag{28}$$

where E_n are the terms of the material stiffness matrix. The other members of the element stiffness matrix can be obtained in a similar way. For additional details the reader is referred to [8]. The derivation of the mass terms and the terms of load vector are also given in [8].

Now we define the special basis functions which enable the approximation of the displacement field without mesh refinement. The simplest such function, which provides for a jump at the endpoint of the Winkler boundary, is the *roof* function. Let ξ_p be a coordinate of the singular point at $\eta = \eta_p = 1$ on the standard domain. (In such a case side 3 of the standard element is subject to the spring boundary conditions. Then the *roof* function is defined by

$$N_r = \phi\left(\xi, \xi_p\right)\frac{1+\eta}{2} \tag{29}$$

where

$$\phi\left(\xi, \xi_p\right) = \begin{cases} \left(1+\xi\right)/\left(1+\xi_p\right) & -1 \le \xi \le \xi_p \\ \left(1-\xi\right)/\left(1-\xi_p\right) & \xi_p \le \xi \le 1 \end{cases}. \tag{30}$$

The definition of the *roof* function for another element edge with spring boundary is similar. Additional functions which have to be included into the finite element space are the ψ-type basis functions which are constructed from the functions of the form which can represent the limiting cases of the analytic solution. These functions have an influence over the domain of the special element only. A cutoff function is needed to ensure that the enforcement of the continuity with other elements is not affected. Therefore, when side 3 is the spring boundary, the definition of the cutoff function is

$$\chi\left(\xi, \xi_p, \eta\right) = F\left(\xi, \xi_p\right)G\left(\eta\right) \tag{31}$$

where

$$F(\xi,\xi_p)=\begin{cases}\left[1-\left(\dfrac{\xi-\xi_p}{\xi_p+1}\right)^2\right]^2 & \xi\le\xi_p\\ \left[1-\left(\dfrac{\xi-\xi_p}{\xi_p-1}\right)^2\right]^2 & \xi>\xi_p\end{cases} \tag{32}$$

and

$$G(\eta)=\left[1-\left(\frac{\eta-1}{2}\right)^2\right]^2. \tag{33}$$

Thus the special basis functions capable of representing the soft spring boundary are

$$\psi_1=\chi r\theta f(\theta) \tag{34}$$

and

$$\psi_2=\chi r\ln r f(\theta) \tag{35}$$

and the essentials of the stiff spring boundary are captured by

$$\psi_j=\chi r^{\alpha/\beta} f\left(\frac{n}{m}\theta\right) \tag{36}$$

where

$$f\left(\frac{n}{m}\theta\right)=\left(a_j\sin\frac{n}{m}\theta+b_j\cos\frac{n}{m}\theta\right) \tag{37}$$

and the best choice of the ratios α/β and n/m are determined based on the lowest eigenvalue of the limiting solution and from numerical experiments which are discussed next.

NUMERICAL EXAMPLES: The performance of the space enrichment method is examined for models of finite two-dimensional bodies. Since no closed form solution exists, the space enrichment method is evaluated against reference solution obtained with Stress Check hp- finite element software. To provide accurate estimates of the potential energy two layers of geometrically graded elements towards the singular point were used [6]. Since the method of approximation minimizes the energy norm of the error, potential energy values are used for assessing the quality of the numerical results. Moreover, construction of ψ-type basis functions also becomes apparent. The ability of the space enrichment method to provide adequate reduction of potential energy is demonstrated by example. The domain and the mesh for the reference solution are shown in Fig. 1. The analysis is performed for a plane strain conditions for a hypothetical material with properties given by $E=1000$ and $\nu=0.3$. A unit normal displacement is prescribed to the boundary AB and symmetry is imposed on boundary AE. The Winkler foundation is characterized by a non dimensional value $R_k=k_s l/E=100$. For the polynomial degree $p=8$ which corresponds to 1600 DOF, the potential energy value obtained with the reference solution is $\Pi_{rs}=363.9945432$. The goal therefore is to obtain an approximation of similar quality with the space enrichment method. The distance between the two solutions is measured by the relative error in the energy norm defined by

$$er \equiv \|e\|_{E(\Omega)} \equiv 100\sqrt{\frac{\left|\Pi(u_{se}) - \Pi(u_{rs})\right|}{\left|\Pi(u_{rs})\right|}}. \qquad (38)$$

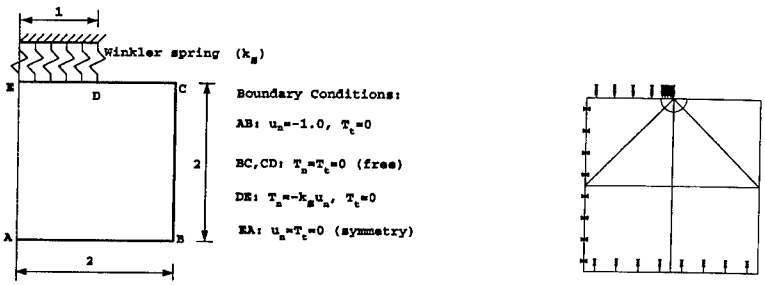


Figure 1. (a) Two-dimensional domain (b) Mesh of the reference solution.

The space enrichment approximation is obtained for a skeleton element formulation based on the tensor product space with $p = 8$, the roof function (29) and functions (34) and (35), which account for the discontinuity in the boundary conditions at the endpoints of the uniform traction loading. The space of the skeleton element is augmented by ψ-type functions, (36). Initially $\alpha/\beta = 1/2$ and n/m are increased from $1/2$ by an increment of $1/2$, thus the singular basis functions are given by

$$\psi_1 = \chi r^{\frac{1}{2}} \sin\frac{1}{2}\theta; \quad \psi_2 = \chi r^{\frac{1}{2}} \cos\frac{1}{2}\theta; \quad \psi_3 = \chi r^{\frac{1}{2}} \sin\theta; \quad \psi_4 = \chi r^{\frac{1}{2}} \cos\theta \ldots \qquad (39)$$

This process is terminated when potential energy value is no longer improving. The results of this experiment are Given in Table 1.

Table 1. The influence of the coefficients *n/m* on the potential energy.

DOF	α/β	n/m	$\Pi(u_{se})$	$er(\%)$
154	1/2	1/2	366.0025877	7.43
158	1/2	1	364.4036265	3.35
162	1/2	3/2	364.2600010	2.70
166	1/2	2	364.2554403	2.68
170	1/2	5/2	364.2539514	2.67
174	1/2	3	364.2532126	2.665
178	1/2	7/2	364.2525488	2.662
182	1/2	4	364.2520090	2.660

This results indicate that only three coefficients in circumferential direction are necessary to provide a significant improvement of the solution. A similar observation was made with respect to other powers of r. Powers and coefficients up to an including α/β=11/2 and n/m=5/2 were examined. With 270 DOF Π_{se} = 363.9995261 which corresponds to a

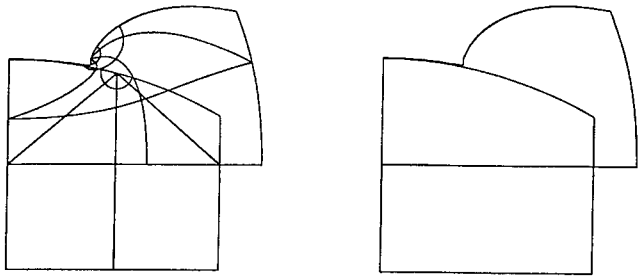
relative error in the energy norm of 0.37% only. Therefore the solution with the space enrichment is of similar quality as the reference solution in terms of potential energy.

In engineering analyses the displacements and stresses are of interest. Moreover, the shape of the contacting boundary may be curved. Therefore the analysis is performed for a domain with circular spring boundary, R=4. The performance of the space enrichment method is examined for two stiffness ratios, $R_k = 1$ and $R_k = 100$, representing soft and stiff boundary conditions respectively. The potential energy values of the reference solutions and the solutions with the space enrichment for 270 DOF, and the relative errors are given in Table 2.

Table 2. Comparison of the space enrichment method with reference solution.

R_k	$\Pi(u_{rs})$	$\Pi(u_{se})$	$er(\%)$
1.	203.3606782	203.3607214	0.05
100.	361.4415500	361.4598269	0.71

Since errors below 1% in energy norm indicate that the accuracy of the finite element solution is generally acceptable, the results obtained for potential energy values are sufficiently accurate for practical purposes. The deformed meshes for the stiff spring ratio, $R_k = 100$, are shown in Fig. 2.



(a) (b)

Figure 2. Deformations corresponding to $R_k = 100$; (a) reference solution (b) space enrichment method.

The deformations are clearly very similar. The singular behavior at the end point of the spring boundary is clearly visible; the displacement changes very substantially over a short interval. The normal component of the stress vector can be estimated indirectly from the boundary conditions, (1), however the tangential component is singular and therefore can not be determined. The comparison of the stresses inside the domain was made for distinct locations; in close vicinity of the singular point and in the points where the solution is smooth. The values of the pointwise stresses are found to be in very good agreement in all locations.

SUMMARY AND CONCLUSIONS: An efficient numerical method for approximation of the solution of elliptic boundary value problem with mixed boundary conditions in two-dimensional elasticity was developed. The finite element method based on the displacement formulation was employed. The finite element space is specially chosen to approximate the solution at the endpoint of the Winkler-type boundary. The construction of the finite element space is based on the fact that the exact solution is smooth everywhere in the domain except in the endpoints of the spring boundary. In these points the solution is not analytic: the nature of the singularity is both material and geometry dependent. Moreover, the exact solution for this class of problems is not known; only the limiting cases are available. Special basis functions have been developed to account for the nature of the exact solution which extend the span of the finite element space. Therefore the method is called the space enrichment method. It is demonstrated that the space enrichment method is very effective for reducing errors of discretization in vicinity of the singular point. The results described in this work provide a basis for further investigation and implementation of the space enrichment method to problems of mechanical contact.

ACKNOWLEDGMENTS: The support of this work by the Air Force Office of Scientific Research under grant No. F49620-93-1-0173 is gratefully acknowledged.

REFERENCES:

[1] A. Signorini, Sopa Alcune Questioni de Elastostatica, Atti della Societa Italiana per el Progressso delle Scienzi, 1933.

[2] G. Duvaut and J.L. Lions, Inequalities in Mechanics and Physics. Springer-Verlag, 1976.

[3]. J.C. Simo, T.A. Laursen, An augmented Lagrangian treatment of contact problems involving friction. Comp. & Struc. 42 No. 1, 97-116, 1992.

[4] I. Páczelt, Solution of elastic contact problems by the finite element displacement method. Acta Techn. Hung. Tom 82 (3-4) 353-375, 1976.

[5] M. Costabel and M. Dauge, A singularly perturbed mixed boundary value problem. Manuscript, 1995.

[6] B. Szabó and I. Babuška, Finite Element Analysis. John Wiley & Sons, Inc., 1991.

[7] N.I. Muskhelishvili, Some Basic Problems of the Mathematical Theory of Elasticity. Noordhoff International Publishing Leyden, 1975.

[8] Y. Volpert, Space Enrichment Methods for the Numerical Solution of Mechanical Contact Problems, Doctoral dissertation Washington University, St. Louis, USA 1996.

DYNAMIC SPHERICAL CAVITY EXPANSION OF THERMAL-VISCOPLASTIC MATERIAL

Zheng Jian Lim Chweeteck

Department of Mechanical and Production Engineering
National University of Singapore
10 Kent Ridge Crescent
Singapore 119260

Abstract: A void growth relation is developed to study the inertial and thermal effects on dynamic growth of voids in ductile materials. Numerical analysis indicates that both inertial and thermal effects predominate behavior of void growth at high strain rate.

Key Words: Damage evolution; Inertial effect; Thermal effect; Numerical analysis; Dynamic void growth; Dynamic fracture

INTRODUCTION: Dynamic damage or fracture in ductile solids is a complex phenomenon, due to a number of different physical effects involved. A number of experimental observations[1] show that dynamic ductile fracture is a consequence of the nucleation, growth, and coalescence of voids in a triaxial stress field. The spall process is found to be temperature sensitive. The spall strength is also found to be dependent on temperature which decreases as temperature increases[2]. Both the yield strength Y and viscosity decrease when temperature increases[3].

The influence of thermal and inertial effects on void growth at high strain rate in a thermal-viscoplastic solid is investigated by means of a theoretical model presented in this paper. The numerical analysis of the theoretical model performed indicates that thermal effect greatly influences void growth at high strain rates. The inertial effect with temperature dependence is also found to be important for the dynamic growth of void, which appears to resist void growth, especially in extremely high strain rates. Besides, numerical analysis shows that material viscosity is also one of the important factors controlling void growth.

VOID GROWTH MODEL: We assume that the porous material is consist of a suspension of pores in a matrix of ductile solid which is subjected to an external stress Σ_{ij}, and that the porous material is statistically homogeneous and isotropic so that it can be effectively modeled by a homogeneous isotropic solid material. In order to simplify theoretical analysis, we also assume that the matrix material is incompressible during void growth and that the void remains spherical all the time. These assumptions simplify the theoretical analysis and allow us to obtain exact relations for the void

growth. Considering a spherical hole of the matrix material of inner radius a and outer radius b(Fig.1), distention α is defined as

$$\alpha = \frac{b^3}{b^3 - a^3} \tag{1}$$

Our purpose is to investigate the response of this hollow sphere when subjected to time-dependent external stress and zero internal pressure, and attempt to obtain the relation between an external stress Σ_{ij} and distention α.

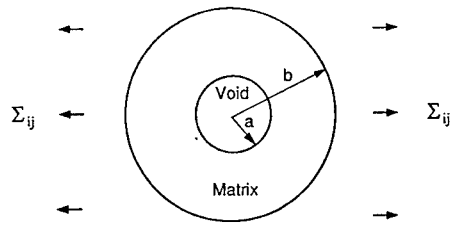


Fig.1 A spherical model of void in the material

Take the matrix material and void as a system, the work done by external applied stress Σ_{ij} is equal to the change in the system energy

$$W = \Delta E_k + \Delta E_i \tag{2}$$

where ΔE_k and ΔE_i denote the changes in kinetic energy and internal energy of the system, respectively. W denotes work done by external stress Σ_{ij}. Since the change in porosity during initial elastic and elastic-plastic phases is quite small[4], we will consider fully-plastic deformation in the solid around the void directly. A general spherical symmetric motion gives the following solutions[5]

$$r^3 = r_0^3 - B(t), \qquad B(t) = \alpha_0^3 \frac{\alpha_0 - \alpha}{\alpha - 1} \tag{3}$$

$$\frac{B(t)}{\alpha_0^3} = \frac{\alpha_0 - \alpha}{\alpha - 1}, \qquad \frac{B(t)}{b_0^3} = \frac{\alpha_0 - \alpha}{\alpha} \tag{4}$$

where r and r_0 denote Eulerian and Lagrangian spherical polar coordinates, respectively. The subscript 0 used in this work denotes at initial stage. B(t) is a function related to the rate of void growth.

Consider ΔE_k, ΔE_i and W, respectively. ΔE_k is given by

$$\Delta E_k = E_k(\alpha) - E_k(\alpha_0) \tag{5}$$

$$E_k(\alpha) = \left[\frac{4\pi a_0^2}{9(\alpha_0 - 1)}\right]\frac{\alpha_0^2 \rho_s}{2(\alpha_0 - 1)}\left(\frac{\alpha_0 - 1}{\alpha - 1}\right)^{1/3}\left[1 - \left[\frac{\alpha - 1}{\alpha}\right]^{1/3}\right]\dot{\alpha}^2 \tag{6}$$

where ρ_s is the density of the matrix material.
In order to treat thermal effect in dynamic void growth, a simple form of linear thermal softening is chosen in the constitutive relation in the matrix material[6]

$$\sigma_e = \left(Y_0 + \eta\dot{\varepsilon}_e^p\right)\left(1 - \frac{T}{T_m}\right) \qquad T \le T_m \tag{7}$$

where T is temperature and T_m is melting temperature of the matrix material. σ_e and $\dot{\varepsilon}_e^p$ denote the equivalent stress and equivalent plastic strain rate in the matrix material, respectively. Y_0 is yield strength of matrix material and η is material viscosity. Since we are dealing with high strain rate, the heat generated by plastic deformation cannot dissipate itself, the temperature then approximately satisfies [7]

$$\frac{\partial T}{\partial t} = \frac{\kappa}{\rho_s C_v}\sigma_e \frac{\partial \varepsilon_e^p}{\partial t} \tag{8}$$

where κ is a constant which is taken as 0.9[8]. C_v is heat capacity of the matrix material. Since we assume a plastic deformation process with spherical symmetry, the equivalent plastic strain ε_e^p is given by Johnson and Mellor[9] as

$$\varepsilon_e^p = 2\ln\frac{r}{r_0} \tag{9}$$

The change of internal energy in the system is

$$\Delta E_i = \frac{1}{\rho_s} r \int_a^b \left[\int_0^{\varepsilon_e^p} \sigma_e d\varepsilon^p\right] 4\pi\rho_s r^2 dr \tag{10}$$

With the help of Eqs.(3), (4) and (9), Eq.(10) can be expressed as

$$\Delta E_i = \left[\frac{4\pi a_0^2}{9(\alpha_0 - 1)}\right]\left[F_1(\alpha) + F_2(\alpha)\right]\dot{\alpha} \tag{11}$$

where

$$F_1(\alpha) = 2Y_0\left(1 - \frac{T_0}{T_m}\right)(\alpha - \alpha_0)\ln\frac{\alpha}{\alpha - 1} \tag{12}$$

$$F_2(\alpha)=\frac{2}{3}\left(1-\frac{T_0}{T_m}\right)\left(1+\frac{2}{3}\frac{Y_0\kappa}{T_m\rho_s C_v}\right)\eta\ln\frac{\alpha}{\alpha-1} \tag{13}$$

Functions $F_1(\alpha)$ and $F_2(\alpha)$ denote the influence on yield stress of matrix material and material viscosity with an increment in internal energy, respectively. The work W done by external stress is expressed as

$$W=\frac{3}{4}\pi\frac{\alpha_0^3}{\alpha_0-1}\int_{\alpha_0}^{\alpha}\left[\frac{2}{3}\Sigma_e(\alpha')-P(\alpha')\right]d\alpha' \tag{14}$$

where Σ_e and P denote the macroscopic equivalent and hydrostatic stresses, respectively. We assume Σ_e and P to be dependent on distention α, that is $\Sigma_e=\Sigma_e(\alpha)$, $P=P(\alpha)$.

Substituting the expressions of ΔE_k, ΔE_i and W into Eq.(2) gives

$$\dot{\alpha}=\frac{1}{F_3(\alpha)}\left\{-F_2(\alpha)+\sqrt{\left[F_2(\alpha)\right]^2-4F_3(\alpha)F_4(\alpha)}\right\} \tag{15}$$

with

$$F_3(\alpha)=\frac{\rho_s a_0^2}{2(\alpha_0-1)}\left(\frac{\alpha_0-1}{\alpha-1}\right)^{1/3}\left[1-\left(\frac{\alpha-1}{\alpha}\right)^{1/3}\right] \tag{16}$$

$$F_4(\alpha)=F_i(\alpha)-3\int_{\alpha_0}^{\alpha}\left[\frac{2}{3}\Sigma_e(\alpha')-P(\alpha')\right]d\alpha'-F_3(\alpha_0)\dot{\alpha}_0^2 \tag{17}$$

Equation (15) is the evolution equation for dynamic growth of void with temperature dependence and is dependent on inertial effect, thermal effect, material strain rate sensitivity and external deviatoric stress. If inertial effect is ignored, $\dot{\alpha}$ is reduced to

$$\dot{\alpha}=-\left[F_4(\alpha)+F_3(\alpha_0)\dot{\alpha}_0^2\right]/F_2(\alpha) \tag{18}$$

For void growth, $\dot{\alpha}\geq 0$ and from Eq.(15), the threshold stress for void growth $\Sigma_{crit}(\alpha)$ is

$$\Sigma_{crit}(\alpha)=\frac{2}{3}Y_0\left(1-\frac{T_0}{T_m}\right)\left(\ln\frac{\alpha}{\alpha-1}-\frac{1}{\alpha}\frac{\alpha_0-1}{\alpha-1}\right) \tag{19}$$

where $\Sigma(\alpha)=\frac{2}{3}\Sigma_e(\alpha)-P(\alpha)$ represents the sum of the applied macroscopic hydrostatic and deviatoric stresses.

NUMERICAL ANALYSES: A copperlike material, with ρ= 8.92g/cm^3, T_m=1356K, C_v=385J/kg K, a_0=1.0003, is selected for numerical analysis. In Fig.2 we show a comparison of the threshold stress versus distention α calculated using the temperature-dependent relation[Eq.(19)] for T_0=373K. The experimentally measured results of spall strength for copper are in the range of 1.0-2.5 GPa[11]. The computed threshold stress when considering thermal effect is lower than that without considering thermal effect. This is especially so for initial threshold stress $\Sigma_{crit}(\alpha_0)$. $\Sigma_{crit}(\alpha_0)$ is 1.78 GPa when excluding thermal effect, but is 1.23 GPa when including thermal effect and they are all in agreement with experimental data. The influence of initial temperature T_0 on initial threshold stress $\Sigma_{crit}(\alpha_0)$ is shown in Fig.3. The initial threshold is noted to decrease as the initial temperature increases. Fig.4 shows the rate of distention $\dot{\alpha}$ versus distention α for different initial temperatures and with material viscosity η=0.7GPa μs.

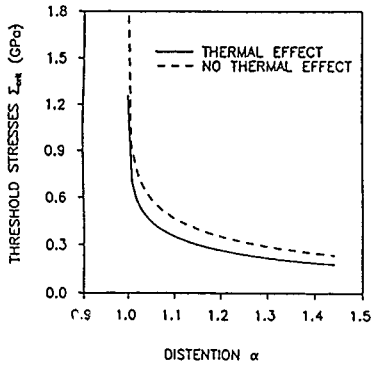


Fig.2 Curves of critical threshold stress Σ_{crit} vs distention α

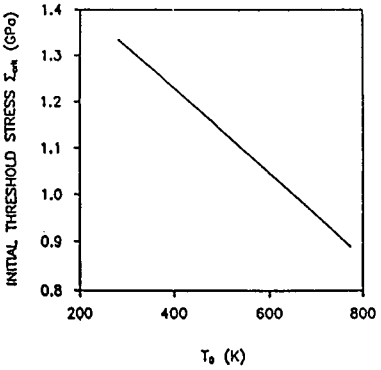


Fig.3 Influence of initial temperature T_0 on initial threshold stress Σ_{ij}

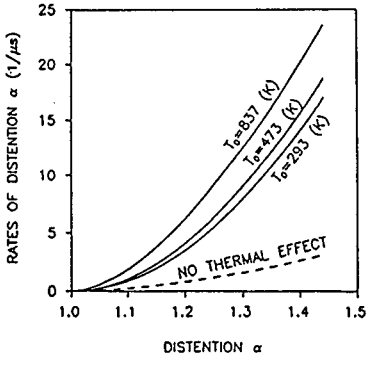


Fig.4 Curves of rate of distention α vs $\dot{\alpha}$ distention α for different initial temperature

The influence of inertial effect for different material viscosity at an initial temperature T_0=373K are illustrated in Fig.5. It is evident that the inertial effect increases with decrease in material viscosity, i.e., as strain rate increases, inertial effect also increases. Moreover, it is also shown that the inertial effect appears to resist void growth.

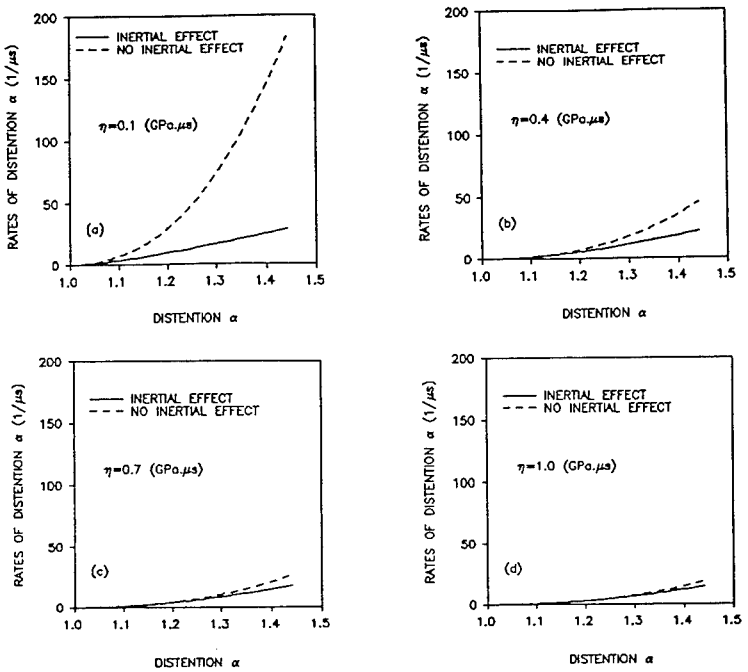


Fig.5 Influence of inertial effect on void growth(Eqs.15 and 18)

CONCLUSIONS: A model of void dynamic growth in ductile solids, which incorporates thermal and inertial effects and the material rate sensitivity is proposed. An expression of the threshold stress for dynamic void growth is also obtained and is found to be dependent on porosity of the materials and initial temperature. Theoretical analysis presented in this paper suggests that the thermal effect on void growth is significant at high strain rates and it increases the rate of void growth. The threshold stress which includes thermal effect for void growth is smaller than that without considering thermal effect. Numerical calculations show that the inertial effect appears to resist void growth and the higher the strain rate, the greater the inertial effect. Thus the dynamic growth of voids is highly sensitive to the strain rate. Therefore the inertial effect should not be neglected in the whole process of spall fracture.

REFERENCES

[1] Curran, D. R., Seaman L., and. Shockey D. A., 1987, Dynamic failure of solids, Physics Report Vol.7, p254.

[2]Grady, D. E., The spall strength of condensed matter, 1988, Journal of Mechanics and Physics of Solids, Vol.36, p353.

[3] Curry, D. A. and Knott, J. F., 1979, Effect of microstructure on cleavage fracture toughness of quenched and tempered steel, Metall. Sci., Vol.13, p341.

[4] Johnson, J. N., Dynamic fracture and spallation in ductile solids, 1981, Journal of Applied Physics, Vol.52, p2812-2825.

[5] Carroll, M. M. and Holt, A. C, Static and dynamic pore-collapse relations for ductile porous materials, 1972, Journal of Applied Physics, Vol.43, p626.

[6] Cortes, R., The growth of microvoids under intense dynamic loading, 1992, International Journal of Solids and Structure, Vol.29, p1339-1350.

[7] Tvergaard, V. and Needleman, A., Effect of material rate sensitivity on failure modes in the Charpy V-Notch test, 1986, Journal of Mechanics and Physics of Solids., Vol.34, p213.

[8] Taylor, G. J. and Quinney, H., 1934, Proc. R. Soc. London, Vol.143A, pp307-326.

[9] Johnson, W. and Mellor, P. B., 1973, Engineering Plasticity(Van Nostrand Reinhold).

[10] Davison, L. and Graham, R. A., Shock compression of solids, 1979, Physics Report, Vol.55, p257.

MFPT

DATA COLLECTION FOR VIBRATION DIAGNOSTICS

Chair: James W. Taylor
Machinery Management Services

A COMMON SENSE APPROACH TO DATA COLLECTING: DATA COLLECTING IS MORE THAN PUTTING A PROBE ON A POINT AND PUSHING A BUTTON

Jim Flowers
CITGO Petroleum Corporation
P.O. Box 1562
Lake Charles, Louisiana 70602

Author Background: Jim Flowers has worked as an outside machinist at CITGO Petroleum Corporation in Lake Charles, LA, since 1985 and has been in the vibration field for five years. Flowers also has thirteen years experience with Texaco as a journeyman machinist. He owned and operated a machined shop for several years. Flowers belongs to the Vibration Institute and has obtained several certifications, including CSI Vibration Analyst Level I, Vibration Institute Specialist I, and Vibration Analyst Level II.

Abstract: In a competitive global market, maximizing mean time between failures (MTBF) while holding down maintenance costs is vital. To this end, the vibration data collector is a valuable resource. However, he can become even more valuable to the company that will 1) use rotating equipment craftspeople (machinists) to collect vibration data, 2) use these craftsmen's experience and knowledge along with the modern technology available to help find root causes of failures, and 3) establish formal communications between these data collectors and others who are involved with making decisions about equipment.

Key Words: Jackbolt; MTBF; resonance; thrust bearing; vanepass

Rationale: Such an integration of technology, common sense, and experience should optimize root-cause analysis, resulting in higher plant reliability, lower maintenance costs, and an improved relationship between craftsman and equipment.

The following example illustrates the wisdom of this approach: A data collector analyzing a 200 hp motor observed a vertical vibration reading of .283 ips and a horizontal vibration reading of .035 ips, nearly an 8:1 ratio. The vertical, horizontal and axial spectra all showed a high 1X rpm and harmonics. A resonance problem was suspected mainly because of the ratio of the horizontal and vertical vibration readings. The data collector noticed that the motor was bolted to a two-inch transition piece that overhung the original base motor supports by six and one-half inches. A vertical reading was taken on the transition piece and showed a vibration of .272 ips (Fig 1). A pair of machinist planner jacks was placed under the transition piece and snugged up while the motor's vibration was monitored. At first the vibration jumped to .78 ips when the planer jacks were snugged, but it came down immedi-

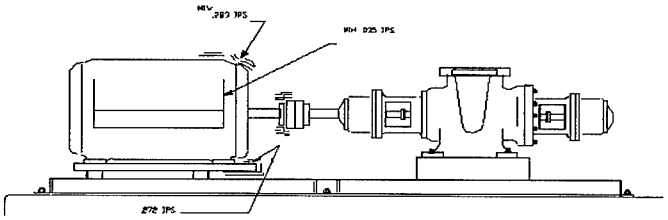


Fig. 1

ately to .035 ips when tightened one quarter turn. A mechanical work request was turned in to modify the base to stiffen the transition piece. As shown in Figure 2 the vertical vibration has not changed much in the last two years since the base was modified. There was also a significant decrease in the 1Xrpm spike and harmonics. There was not a big change in the horizontal readings (Fig 3), but the readings were consistent and not erratic, which is common with a resonance problem.

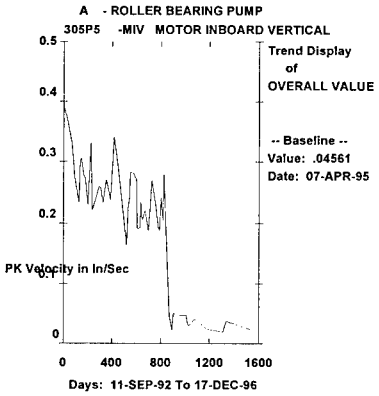


Fig. 2

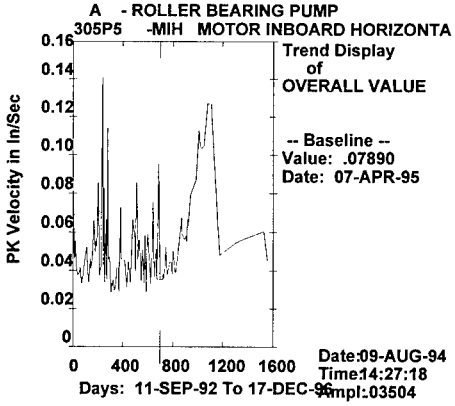


Fig. 3

Pipe Stress: One advantage a machinist offers as a data collector is the knowledge that pipe stress is just as critical as the repair, balance, and alignment (Fig. 4). Sometimes foremen of other crafts not related to rotating equipment (pipefitters, welders, carpenters, etc.) have been put in charge of an alignment job where the equipment had more pipe stress than what was acceptable (normally .002 in.). Some of these foremen believed that excessive pipe stress on a pump could be relieved by realigning the motor to the pump. Actually all that was accomplished was aligning the equipment; the pipe stress had not changed. The base bolts were keeping the pump from moving, and the pipe stress was actually distorting the case and all the parts on the inside, such as the seals, bearings

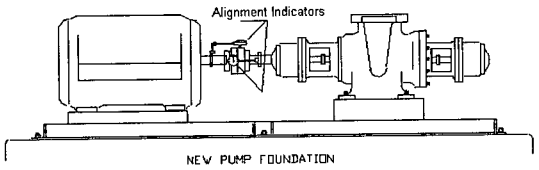


Fig. 4

and wear rings, as well as putting excessive stress on the case and flanges. The pipe stress could cause the seal sleeve to rub the throat bushing, cock the bearings, or cause the impeller to rub the wear rings. Pipe stress is one of the root causes that is very hard to diagnose on a piece of critical equipment while it is in service.

Bearing housing to sleeve bearing crush: Another advantage a machinist brings to data collecting is a grasp of the importance of having the correct *crush* (clearance between the OD of the sleeve bearing and the ID of the case) on a sleeve bearing (Fig. 5), whether on a pump, turbine, blower, or gearbox. The machinist understands that if the crush is too loose, the bearing can flop around, causing excessive wear. Also, crush can prevent the oil rings from lubricating the equipment properly. On the other hand, a too-tight bearing crush could deform the bearing, closing the clearances on the ID of the sleeve bearing and causing a premature failure.

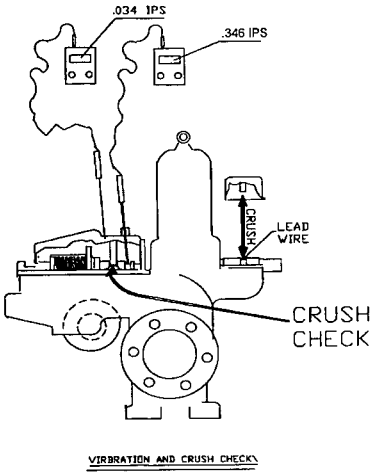


Fig. 5

Few people seem willing to discuss sleeve bearings or criteria for running vibration routes on small general purpose turbines with sleeve bearings, to detect a sleeve bearing before it fails, particularly the bronze type and the metal/babbitt lined, both of which use oil rings or oil mist for lubrication. Often when a sleeve bearing fails, the vibration levels do not always increase, and when they do it is not much (maybe .05 ips). In approximately 50 percent of the sleeve bearing failures encountered, the vibration actually decreased. The assumption then is that when either the sleeve bearing or shaft galls, it closes the bearing clearance which would decrease the peak to peak displacement, accounting for the decrease in vibration.

The advice given by many turbine manufacturers (OEM) as well as some highly respected consultants in diagnosing problem sleeve bearings is "get concerned when the overall velocity readings reach .30 ips or install non-contact displacement probes." Since it is not economically feasible in most cases to install non-contact probes on these general purpose turbines, the overall velocity readings must be taken. However, the majority of the turbines never reach .30 ips before they lose a bearing and have to be repaired, and most of the time the shaft has to replaced.

A one-week informal survey on 100 of these general-purpose turbines was conducted while running vibration routes. The average overall velocity readings of these 100 turbines was .05 ips, a considerable distance from .30 ips. The current standard of .30 ips may be too high an alarm level. One solution may be to notice any readings that increase or decrease by one half of the previous reading and do more checking. It is much cheaper to change a sleeve bearing in the field than to pull either the rotating element or the complete turbine for repairs.

An experienced data collector may also be able to diagnose sleeve bearing problems another way. If the bearing crush is correct, there is not much difference in the vertical vibration readings on the bearing cover or vertical readings on the sleeve bearing itself. Machinists set the majority of the bearing crush at metal/metal to .0005 in. loose. One recommendation in diagnosing sleeve bearings is to include the readings on top of the sleeve bearing itself as a reference when taking base line readings after an overhaul or new installation.

In one actual episode, the reading on top of a sleeve bearing was compared to the vertical reading on top of the bearing housing. The first 35 or 40 that were checked showed the readings were within .05 ips of each other. Then one showed vibration readings of .034 ips (Fig 5) on the bearing cover but .346 ips on top of the sleeve bearing. The reading on the sleeve bearing was 10 times the bearing housing reading. Further investigation showed that the sleeve bearing could be moved from one side to the other. A maintenance work request (MWR) was sent in for a crush check on the sleeve bearing during the next outage. When the bearing crush was checked, it was .0045 in. loose to the case. The bearing cover had .004 in. taken off the split line in the shop, which resulted in the correct crush. The turbine is still running two years later.

A second turbine found with a loose sleeve bearing showed vibration readings of .075 ips on top of the bearing housing and .647 ips on top of the sleeve bearing. The machinist found the bearing crush to be .005 in. loose. The bearing cover was then ground to the correct crush. Understanding how a turbine works enables employees to help save time and money by allowing operations time to schedule the outage for repairs instead of the turbine failing at a time that could affect production or cause secondary damage. The machinist's experience provides an insight some data collectors may not have.

Thrust: The bearing thrust on the majority of overhung and split case pumps, both with mechanical seals, is set at .001 in. to .0025 in., depending upon the operating temperature of the equipment. This knowledge helped to detect and repair several split case pumps where the thrust bearing nuts had loosened before they caused any secondary damage or lost production.

Something wrong in the thrust bearing end of the pump could cause a shaft to thrust axially. The problem could be one of the following: 1) the bearing nut or retainer had loosened, 2) the thrust bearings were bad, or 3) the shaft was broken. Under normal conditions if the flow for that particular pump was correct, there should be little or no thrust, but thrust set at

.0025 in. cannot be seen with the naked eye. Therefore, the pump would not thrust unless something was wrong with the thrust bearings.

Third legs on overhung pumps: Another advantage the rotating equipment craftsperson (machinist) has as a data collector is the understanding that the third leg on a overhung pump is only for stability and not for alignment. The third leg is the last thing that is tightened up after the alignment is completed. The machinist knows to use dial indicators to check movement while tightening the third leg. If a hot alignment has to be done, the third leg needs to be loose while the equipment is heating up.

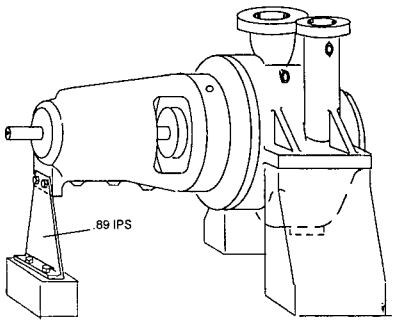


Fig. 6

Third legs come in all shapes and sizes, and sometimes they are the root cause of a pump's premature failure. On several occasions the third leg has set up a resonance, resulting in vibration readings on the third leg as high as .89 ips to .95 ips (Fig 6). These were problem pumps that had to be repaired on the average of every two to three months from 1986 until 1992, mainly to replace the thrust bearings. The original third leg was a triangular shape that measured 22 in. tall, 5 in. across the top, and 12 in. across the bottom, made out of 3/16 in plate. One suggestion was made to the reliability engineer that a third leg be constructed out of 3/8 in. plate (after it was determined that it would not affect the thermal growth) to see if that would help decrease the vibration and help the mean time between failure on this pump. After installing the new third leg, the pump ran two years before repairs were required on a seal which failed due to contamination, but the bearings were still fine .

Also, all third legs are not mounted on the pump base in the same manner. Some use a jackbolt and jamnut (Fig 7), some third legs just rest on the pump base (Fig 8), but the best are the ones that are bolted to the pump base (Fig. 6). Many OEM's agree that it is better to bolt the third leg to the pump base.

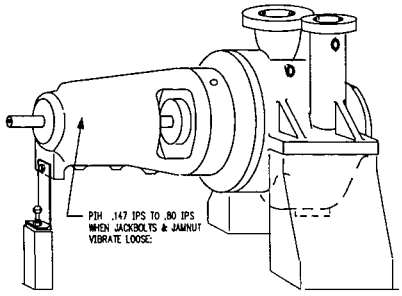


Fig. 7

Over a two-and-one-half-year period, a pump that used the jackbolt and jamnut method had several problems. The jackbolt would vibrate loose and the overall vibration level would rise from .147 ips to over .80 ips. It would require the combined efforts of the area reliability engineer (to monitor the equipment while the machinist adjusted the jackbolt and jamnut), one machinist (to adjust the jamnut and jackbolt), and one or two operators standing by just in case they had to shut the equipment down. In this case of adjusting the jackbolt with the

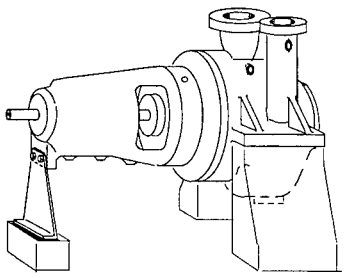


Fig. 8

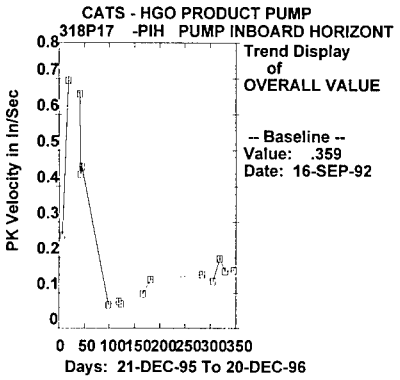


Fig. 9

equipment running, it is possible to adjust it too much and cause the bearings to heat up due to misalignment. This process would normally take about an hour to accomplish and would have to be done sometimes as often as once a week. The last time this pump was pulled, it was suggested that the pump base be drilled and tapped to eliminate the jackbolt and jamnut, and the third leg could be bolted solid to the base. At first the maintenance foreman and reliability engineer objected because they said by using the jackbolt and jamnut the pump could float when the unit's flow rate and temperature changed, which ranged between 400 and 550 degrees F. The machinist explained that if the pump floated with the flow and temperature change the alignment was floating also. They finally agreed to drill and tap the base and bolt the third leg solid. Not only has the vibration decreased and remained at the lower rate (Fig. 9), but at least 18 employee-hours per month were saved. Also, some high-frequency spikes in the pump's spectrum that could not previously be accounted for completely disappeared.

Communication: Step 1 for a data collector or analyst going into a unit to run a vibration route is to ask the operator if there is any piece of equipment that they would like checked (since the data collector is only around the equipment every week or two and the operators are around it seven days a week). The operators know to tell the data collector of any changes they have noticed because if there have not been any operational changes, it could very well mean catching a potential problem before it causes an unscheduled outage. The data collector should then let the operators know if any

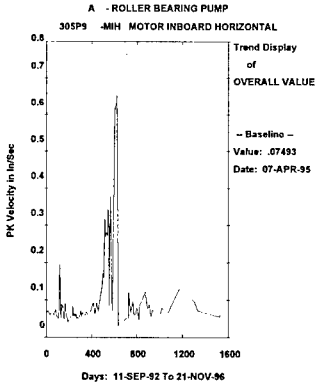


Fig. 10

problems are found, and whether they need to be concerned with it or just keep an eye on it for any change that might take place.

There have been several times that the equipment did in fact have to be shut down because the problem worsened. A prime example of this is a 100 hp motor that had been monitored for three years (Fig. 10). When readings were taken one day, the overall velocity readings were only .075 ips. A second set of readings showed .45 ips, six times as high as the first set of readings. A third set of readings showed a level back down to .075 ips. When timed, the readings would fluctuate from .075 ips to .45 ips every five seconds. Since there was no spare motor in the warehouse, one was ordered. The vibration readings went from an erratic reading to a constant vibration reading over the next several weeks, as high as .698 ips. Approximately six weeks after the erratic readings were discovered, operations had to shut the motor and pump down and replace the motor. By working together and letting everyone involved with the equipment know about the problem, we were able to safely monitor the equipment, get another motor ordered, and save a lot of unnecessary downtime.

Decrease the vibration and increase the life of the equipment: It has been shown that if the vibration is decreased by one-half, the life of the equipment can be extended five times. If the vibration can be decreased by more than one-half, the life of the equipment can be increased by up to fifteen or twenty times. It has been shown that a .2 ips -.4 ips increase in vibration can mean a 70 % decrease in bearing life. A good example to confirm that the theory works is the pump with the resonance problem on the third leg mentioned earlier (Fig 6) that was being repaired on the average of every two to three months: After a new third leg was installed, it performed for over two years. Repairing such things as misalignment, replacement of bearings on double-ended pumps and turbines before they cause secondary damage, and stiffening of third legs that have a high vibration due to a resonance problem can help increase the life of the equipment. Although an outage may have to be scheduled at first, in the long run it could extend the life of the equipment and lower maintenance costs without affecting production due to an unscheduled outage.

The knowledge gained through the additional training companies provide has paid for itself many times over. After listening to noted speakers such Art Crawford, Ron L. Eshleman, Charlie Jackson, and Ed Nelson, the data collector learned a little bit more to help in the war against *unwanted, destructive vibrations*.

For example, Art Crawford's report on resonance resulted in the detection and repair of the resonance problem on the third leg mentioned earlier [1]. Charlie Jackson's lecture on diagnosing problems such as misalignment, imbalance, bad or locked-up couplings, and resonance using phase angles emphasized the importance of taking base line phase readings along with the other vibration baseline readings [2]. One helpful hint in becoming a better data collector came from Ken Jackson at a user's conference. He stated that even though the time waveform is a complex arrangement and could be used to find a problem that would not show up in a spectrum, when used in routine data collecting the data collector can tell immediately the severity of the problem. He said, "If it takes longer than 15 seconds to diagnose the severity of the problem from the waveform, time is just being wasted.[3]

Vanepass: Countless pumps have been torn down to find the case and impeller badly eroded, sometimes to the point of having to be replaced if one was available. The problem was evidenced by the primary vibration frequency showing up at 4 to 8 times running speed. Also a lot of ground noise was noticed in the spectrum raised off the floor or baseline. Then the correlation became clear: The vanepass was showing up at the number of vanes of the impeller times the running speed, and was caused by either the pump cavitating or recirculation due to either an eroded impeller or a too-large gap between the shroud of the impeller and the cut water.

An article by Ed Nelson in *Pumps and Systems* tells about trimming just the vanes of the impeller and not the shroud, leaving the correct gap between the OD of the impeller and the cut water in the pump case. Mr. Nelson stated that this could cut down on the erosion of the pump by eliminating much of the recirculation and extending the life of the equipment because of the decrease in vibration [4]. After discussing this with some engineers who also had read Mr. Nelson's article, we agreed that trimming the vanes and not the shroud could increase MTBF. Two forms have been developed to use when the pump is disassembled (See Figures 12 and 13).

The common sense approach to data collecting: Approximately half of the problems the data collector discovered were due to a sharp increase or decrease in the vibration levels instead of just a high overall reading. After noticing the sharp increase, the data collector looked at the spectrum and relevant parameters to see if the increase was related to a mechanical problem or an operational problem. Then the time waveform was evaluated to determine the severity of the problem. The following case histories are common problems found in industry today. In each case 1) the overall vibration levels never reached a level that most severity charts would say to get concerned about, and 2) the problem was caught due to a sharp increase.

Case history #1: A 50 hp turbine's vibration increased from .048 ips to .114 ips in four days (Fig. 11). As mentioned earlier, the severity charts said the readings were fair to slightly rough, but experience aroused suspicion of that much increase in just four days, even though the overall vibration level was not high. The operator was asked if he had noticed any difference in the turbine, which he had not, but he stated the pump's flow was cut back because an exchanger was being worked on. Operations preferred to run this particular turbine on the main steam valve instead of using the governor to control its speed. The operator was told to open the steam valve a little to determine if that would help the vibration.

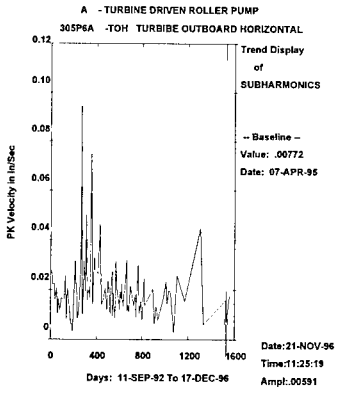


Fig. 11

CITGO Petroleum Corporation
Lake Charles Operation
INTEROFFICE LETTER

Impeller Trim Performance Form

TO: DATE:

FROM: Phone: Pager:

CMMS ID:_____

MAX. DIA. IMPELLER_____ (cmms)

	BEFORE	AFTER
TDC FLOWRATE (BPD/GPM) (CIRCLE ONE)		*
DISCHARGE PRESSURE (PSIG) +		
SUCTION PRESSURE (PSIG/PSIA) + (CIRCLE ONE)		
PUMPING TEMPERATURE (°F)		
SPEED (RPM) (REQUIRED FOR TURBINES)		
CURRENT (AMPS) (IF AVAILABLE)		
IMPELLER DIA. (IN) (SHOP VERIFY)		

FLOWMETER LOCATION: ☐ UPSTREAM ☐ DOWNSTREAM OF PUMP

COPY: AREA RELIABILITY ENGINEER_____

* AFTER IMPELLER TRIM MATCH "BEFORE" FLOWRATES FOR TEST

+ USE CALIBRATED GAUGES OR USE ONE GAUGE FOR BOTH SUCTION AND DISCHARGE READINGS

Figure 12

PUMP IMPELLER TRIM DATA SHEET

CMMS ID:_____ W.O. #:_____ DATE:_____

NO. OF PUMP STAGES:_____ NO. OF IMPELLER VANES:_____

PUMP STAGE NO.:_____ INLET EYE DIA.:_____ Deye

IMPELLER SUCTION: ☐ SINGLE ☐ DOUBLE BALANCE HOLE NO./DIA.:_____

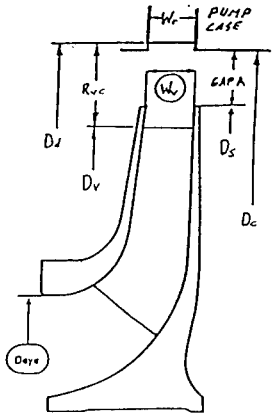
IMPELLER DISCHARGE: ☐ VOLUTE ☐ DIFFUSER VOLUTE/DIFFUSER INLET WIDTH: _____Wr

NO. VOLUTE/DIFFUSER VANES: _____ MAX. DIA. IMPELLER: _____(CMMS)

VOLUTE OR DIFFUSER TIP CONDITION: ☐ O.K. ☐ ERODED
COMMENTS:_____

MEASURE AND RECORD:

	AS-FOUND	AS-LEFT	COMMENTS
VANE DIAMETER -Dv-			NEED DCA TO CHANGE Dv
SHROUD DIAMETER -Ds-			LEAVE AT MAX. DIA. IF POSSIBLE
VANE WIDTH AT SHROUD O.D. -Wv-			
VANE THICKNESS -T-			1/4" MIN. TO 3/8" MAX.
VOLUTE/DIFF.DIA. (ABOVE VANE TIP) -Dd-			
CASE DIA.(ABOVE SHROUD) -Dc-			
VANE RADIUS R=Dv/2			
VOLUTE/DIFF. INLET RADIUS Rd = Dd/2			
IMPELLER SHROUD RADIAL CLEARANCE GAP A = (Dc-Dv)/2			SEE RECOMMENDATIONS BELOW
RADIAL VANE TIP CLEARANCE Rv=Rd-Rv			
GAP B = 100 · Rvr/Rv			SEE RECOMMENDATIONS BELOW

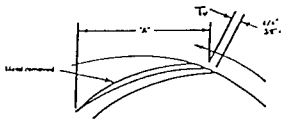


Type pump design	Gap "A"	Gap "B"—Percentage of impeller radius		
		Minimum	Preferred	Maximum
Diffuser	50 mils	4%	6%	12%
Volute	50 mils	6%	10%	12%

"B" = $100(R_d - R_v)/R_v$. R_d = radius of diffuser or volute inlet; R_v = radius of impeller.

NOTE: If the number of impeller vanes and the number of diffuser/volute vanes are both even, the radial gap must be larger by about 4.

Recommended radial gaps for pumps



Length of blend for over filing	
Impeller diameter, in	"A" distance of blend, in
10 & below	1¼
10½ through 18	2¼
19½ through 30	3¼
30½ through 50	4
50 & larger	6

Figure 13

What happened next is everyone's worst nightmare when working around steam turbines. The governor would not actuate the governor valve when the main steam valve was opened, and the trip hammer on the overspeed was trying unsuccessfully to release the trip valve. The spectrum showed this 3600 rpm turbine was running at 4304 rpm and climbing. The operator shut the main steam valve while the unit supervisor and maintenance foreman were notified. The problem was that the trip valve was stuck, the governor valve was stuck, and the governor coupling had slid back on the shaft. It is always stressed to the operators to be careful around steam turbines and let someone know of any change in the way they run, no matter how small.

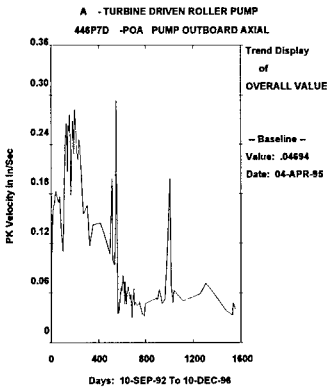


Fig. 14

Case history #2: On March 8, 1994, the axial vibration on the outboard end of the 700 hp Boiler Feed Water Pump had increased from .100 ips to .291 ips (Fig. 14). The horizontal vibration readings never rose over .08 ips, and the vertical readings were never over .075 ips. The data collector spectrum showed a distinct bearing problem had developed on the thrust bearing. A maintenance work request was initiated. The outer races on both thrust bearings had a defect in approximately the same spot. The marks on the bearings showed that the bad spots on both bearings were located at the top of the bearing housing. What was suspected to be the root cause was the packing leaking and spraying a stream of water directly in the vicinity of the bad spots.

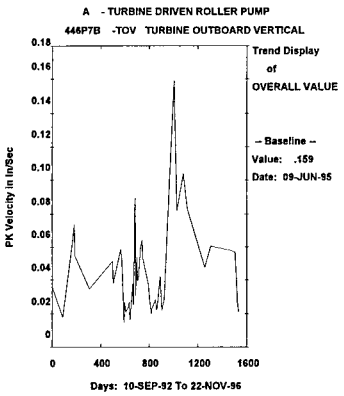


Fig. 15

The machinists were able to change the thrust bearings with no secondary damage to the shaft or bearing housing and put the pump back in service in less than six hours. This pump has been running for about five years with the same bearings. An overhaul on this pump can run as high as $100,000.

Case history #3: As shown in Figure 15, the vertical vibration on the OB (outboard) end of the J-7B boiler feed water turbine did not make any significant changes. In Figure 16, the horizontal reading made an increase and then leveled out, but was

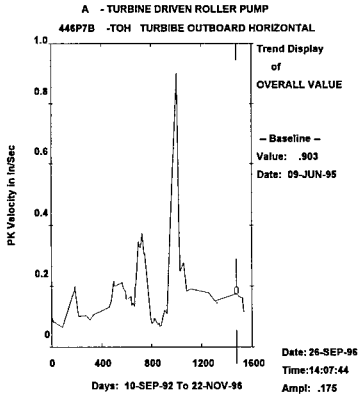


Fig. 16

still not at an alert level. Figure 17 shows the axial reading on the OB end steadily getting higher and then on April 5 making a sharp increase. A maintenance work request was initiated to check the outboard sleeve bearing and also the thrust bearing. The thrust bearing and shaft had started pulling babbitt. The shaft was not hurt and the machinists were able to change the sleeve bearing and thrust bearing and put the equipment back in service with no loss of production and no secondary damage. Waiting until the vibration had reached .30 ips might have been too late. An overhaul on this type of turbine can cost $50,000 or more.

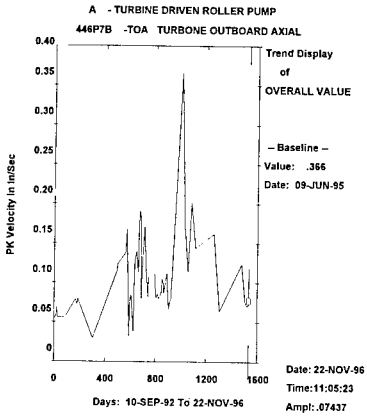


Fig. 17

Case history #4: The axial vibration on the outboard end of the Powerhouse J-5C blower had increased from .045 ips to .125 ips in one week. To ensure accuracy, another set of readings was taken. The second set of vibration readings showed .152 ips. A third set of readings showed .183 ips. (Fig. 18). Also, an oil sample was black, indicating a bearing about to fail. Immediately the maintenance supervisor, unit supervisor, and the area reliability engineer were called.

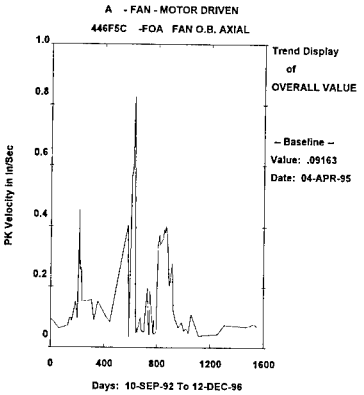


Fig. 18

Immediate flushing of the bearing with fresh oil began. The vibration then increased to .589 ips. After flushing with oil for two and one-half hours, the vibration came down to .07 ips, and the oil stopped turning black. The blower was monitored several times that day and operations kept a close eye on the oil to make sure it was not turning black again. The next day another set of readings was taken, and the vibration had come back down to .045 ips.

The main concern about losing this blower was that there were already two boilers down for repairs, and if this blower went down, one of the main refinery generators would have to be shut down. The blower was shut down four months later. Operations and maintenance were able to schedule the shutdown without affecting refinery operations.

The "data dog" syndrome: In conclusion, using machinists to collect vibration data and analyze it can not only be beneficial to your organization, but it can also instill pride in ownership to the ones collecting the data. The money invested in additional training for

your data collectors on the proper equipment will pay for itself many times over.

The main thing for data collectors to avoid is the "data dog syndrome" or vibration fatigue. It's no fun collecting data eight hours a day, five days a week. Have enough data collectors available so that they can run their routes and have a couple of days to do analysis work, additional training, and work with others in finding the root causes on problem equipment. This makes the data collector's job more fulfilling and frees the engineer to work on more important projects.

References:

1. Crawford, Art. Presentation at the annual meeting of Triplex Chapter of the Vibration Institute; April, 1994.
2. Jackson, Charles. Lecture in Machinery Vibration Analysis I seminar; McNeese State University; Lake Charles, LA; Fall, 1993.
3. Jackson, Ken. Presentation at CSI User's Conference; Nashville, Tennessee; October, 1994.
4. Nelson, Ed. "Addressing Pump Vibrations" (Part I & II). Pumps & Systems (Mar-April 1993).

MFPT

LIFE EXTENSION AND DURABILITY

CoChairs: Michael J. Roemer
Stress Technology, Inc.

Rudolph J. Scavuzzo
University of Akron

REAL-TIME RELIABILITY EVALUATION OF VIBRATING MECHANICAL STRUCTURES

Uwe Kay Rakowsky
University of Wuppertal, Department of Safety Engineering
Safety of Transportation Systems Section, D-42097 Wuppertal, Germany

Dirk Söffker
University of Wuppertal, Department of Safety Engineering
Safety Control Engineering Section, D-42097 Wuppertal, Germany

Abstract: The concept of reliability analysis in which methodical tools are interwoven has remained unchanged for decades. It seems necessary to start on a continuous, so called real-time analysis which combines the methods of control theory with those of reliability theory. A Safety and Reliability Control Engineering concept will be presented, in which signal data are taken from the vibrating mechanical system. These signal data will be converted into reliability characteristics.

Key Words: Failure rate; fuzzy logic; laval rotor; on-line analysis; reliability engineering

INTRODUCTION: Since the late fifties, it has been the engineers' intention in the field of reliability to continuously improve the quality of their methodical tools. Procedures have been developed which make a realistic representation of systems in mathematical models possible and which are also of universal validity for a wide range of application. However, the concept of reliability analysis in which those tools are interwoven has remained unchanged. All procedures were applied during the construction phase of a system. Possible modifications, based on knowledge gained from the analysis, lead to an update of the model. The reliability analysis was finished with putting the system into operation.

It seems to be necessary to leave this narrow period of time and start on a continuous, so called real-time, analysis. This allows the determination of data during the operation of the system, which would possibly correct the results of those analyses done before. A continuously updated analysis allows the optimization of strategies for maintenance, a reduction of expensive down-time, therefore, a higher availability and, most of all, a higher degree of operating safety of the system. To reach these aims, however, it is necessary to combine the methods of control theory with those of reliability theory, since an early discovery of interuptions of operation and impending defects are essential for a predicition concerning the further behaviour of the system.

A new concept will be presented, in which signal data are taken from the vibrating mechanical system. These signal data will be converted into reliability characteristics. After their valuation the operating parameters can be used for control, e.g. the steam input of a turbine is controled depending on the evaluated reliability characteristics. The main emphasis of this paper is the presentation of the transformation of signal data into reliability characteristics. For this transformation eight different ways are suggested, which include the usage of observers, estimators, human respectively artificial experts. Furthermore, interfaces for parameter identification, process characteristics or strain characteristics are intended. Finally, a brief example of application will be presented.

CONVENTIONAL AND EXTENDED RELIABILITY ANALYSIS CONCEPTS: The Integrated Method System (IMS) [4] (figure 1) belongs to the concepts of the development of products and processes. It contains methodical tools for quality control and reliability analysis, which promote the creativity of those working groups enabling the exchange of information and leading to less subjective results through application of valuation techniques. Despite the devision of tasks within the team common aims shall be achieved. This requires both, the willingness of all the people involved to cooperate and the coordination of the single procedures.

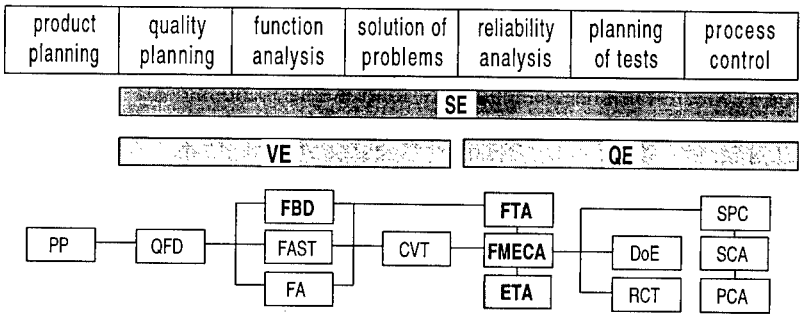


Figure 1 The integrated method system as an example of a conventional concept

The following abbreviations mean: SE - Simultaneous Engineering; VE - Value Engineering; QE - Quality Engineering. The abbreviations for the methodical tools are: PP - Product Planning; QFD - Quality Function Deployment; FAST - Function Analysis System Technique; FA - Function Analysis (hierarchical); FBD - Function Block Diagram; CVT - Creative Value Techniques; FMECA - Failure Modes, Effects, and Criticality Analysis; FTA - Fault Tree Analysis; ETA - Event Tree Analysis; DoE - Design of Experiments; RCT - Reliability Conformance Testing; SPC - Statistical Process Control; SCA - System Capability Analysis; PCA - Process Capability Analysis.

As it can be easily seen, a reliability analysis is carried out and completed before putting the system into operation. The aim of this approach is to add a further element to the process control. Its function is to permanently determine the failure probability of the system during operation, to compare this failure probability to the references given and finally to control the system. The presentation (figure 2) of the approach of *Safety and Reliability Control Engineering* concept (*SRCE*) differs slightly from the representation in the accompaning paper [8].

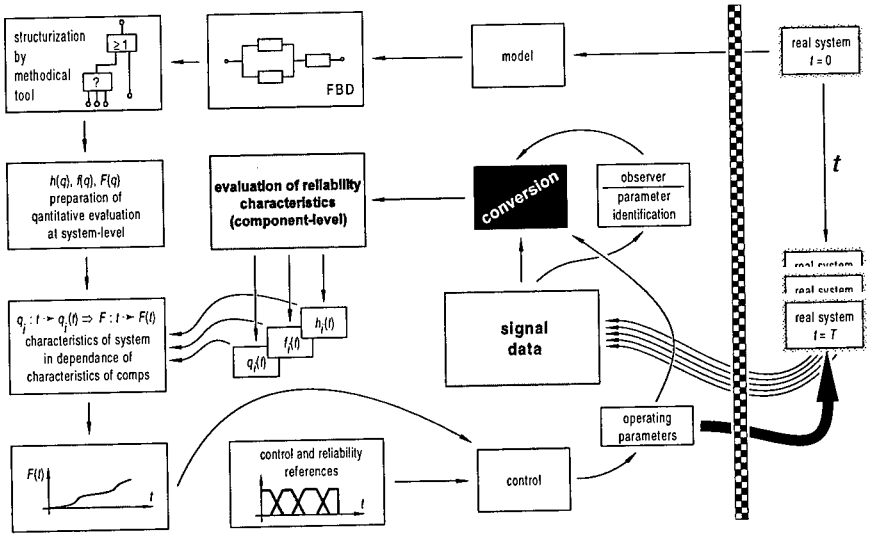


Figure 2 The *SRCE*-concept

Basis of the *SRCE*-concept is to first provide a reliability model of the system before it starts operating. Here the conventional procedures of contemporary Reliability Engineering can be applied. The figures suggest an approach via Function Block Diagrams (FBD), fault tree analysis and its quantitive evaluation. Of course, one could have also used different tools, e.g. Markov-processes, event tree analysis, multi-state coherent system modelling. However, this contribution focuses mainly on the online-transformation (Black Box in figure 2) of signal data taken from the system into reliability characteristics. It will be introduced in the next section and in section 4 the transformation will be illustrated with an example. If the reliability chracteristics are given, they will be compared to the references. According to a control strategy, e.g. "The failure probability of the system should not go beyond 10^{-5}", the operating parameters are evaluated. They then operate on the system to transfer it into the desired state.

CONVERSION OF SIGNAL DATA INTO RELIABILITY CHARACTERISTICS: Centre of the *SRCE*-concept is the actual determination of reliability characteristics from signal data and other information of the system. It is the task of methods of control theory to describe the actual performance of the system. For this, different paths are possible, as presented in figure 3.

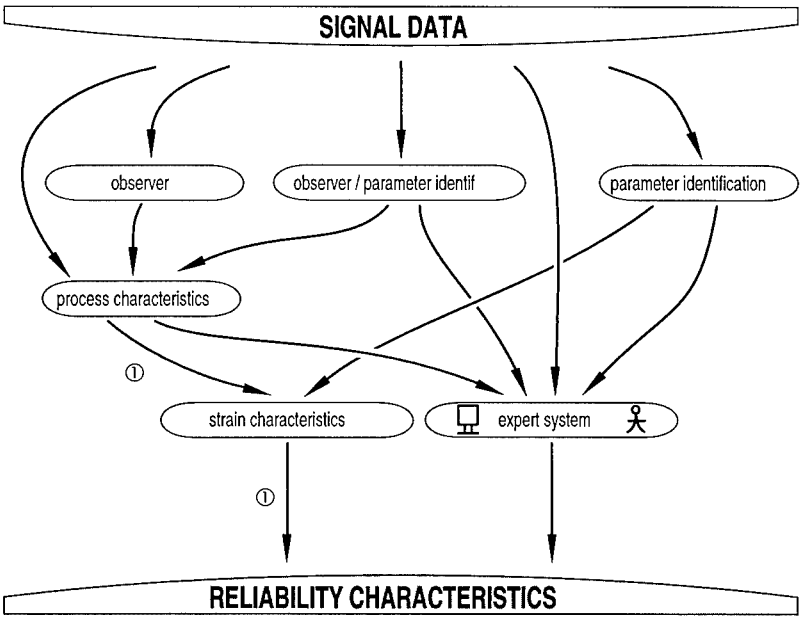


Figure 3 Ways to convert signal data into reliability characteristics

Closer attention will be paid to the observer and parameter identification approaches. The main idea of observer techniques is to build up a mathematical relationship which uses the principal known input-output relations modeled by a system matrix A to use the difference between the real measurements y and the estimated measurements $\hat{y}$ for the estimation $\hat{x}$ of unmeasurable inner states x. This implies that the inner relations of A allow this, using only $y = C\,x$, a part of the system states. This assumption is called observability of the pair (A,C). Related to the interesting area of the application example extended techniques are introduced in [5]. In the ideal case the process variables are formulated as system states to be estimated. To get suitable estimations $\hat{x}$ problem equivalent observer techniques has to be used. In [9] a linear, robust technique is proposed, which allows the robust estimation of states and also of distur-

bances (for example caused by system changes). If $\hat{x}$ does not represent process-coefficients also combinations of $\hat{x}_i$ and elements of A etc. (assumed as constant) can be used for calculation. The next step – the calculation of strain characteristics – can be done in the same way. These steps related to mechanical systems is given in [7]. Parameter estimation approaches also uses the input-output relation modeled by differential equations using a set Θ of parameters a_i, b_i for the input-output relation

$$y(k) + \ldots + a_n\, y(k\text{-}n) = b_0\, u(k) + b_1\, u(k\text{-}1) + \ldots + b_m\, u(k\text{-}m) \qquad (1)$$

as a scalar difference equation or a set of equations between the output y and the input u and k as the actual discrete time step. The parameters a_i, b_i typically do not represent the physical parameters p_i of the system, but if the relation

$$\Theta = f(p_i) \qquad (2)$$

is unique, an inverse transformation f^{-1} gives the physical parameters. The work of Rückwald [6] introduces a fast backtransformation approach. Changes of the physical system are assumed as changes $\Delta\, p_i$, which lead also to changes of the process coefficients a_i, b_i. The idea of FDI using parameter estimation and identification approaches is to observe the parameters (process or physical parameters) using threshold values, some heuristic methods inspired from practice or stochastic decision theory. For solving the estimation process a lot of estimation techniques are available like least squares (LS) or recursive least squares (RLS) techniques. Important is that the assumptions concerning the available measurements φ are

$$det\ E\{\varphi(t)\,\varphi(t)^T\} \neq 0\ . \qquad (3)$$

This includes a high information level concerning the measurements of the system.

Here way ①, signal data ➛ ... ➛ process characteristic ➛ stress characteristic ➛ reliability characteristic, shall be introduced. The conversion of signal data into reliability characteristics occur in discrete time units with $t \in [0,\infty)$; $T_0 = 0$; $T_k \in\ ]0,\infty)$ as well as $K = \{1,2,\ldots,\ell\}$; $k,\kappa \in K$; $\kappa \leq k$. The system failure rate h_k for the k-th interval $[T_{k\text{-}1}\,;\,T_k] \subset [0,\infty)$ shall be determined subsequently. Hereby a constant failure rate λ_k is presupposed to minimalize the expenditure of calculation. If an exact modelling is desired it will be easier to decrease the time interval than to deviate from a constant rate. After the conversion of signal data into m process characteristics depending on time $g_j : t \rightarrow g_j(t)$ with $j \in \{1,2,\ldots,m\}$ the so-called *strain characteristics* are transformed. The strain characteristics $\beta_j : g_j(t) \rightarrow \beta_j(g_j(t))$ are to be understood as filter, which assign different "weights" to the process characteristics depending on the specific properties of the system. This way the graph of β_j generally runs relatively constant within the required operating capacity and ascents outside of it.

A failure rate λ_k for the k-th time interval is determined from all the m values $\beta_j(g_j(t_k))$ for $t_k \in [T_{k-1}\,;\,T_k]$. The manner in which λ_k depends on the strain characteristics is expressed by means of fuzzy logic. There has been quite a few publications within the last years on the topic of fuzzy logic [1, 2, 3, 10], therefore the decision making will only be briefly presented in an example without any introduction to fuzzy logic. If m strain characteristics are given, formulated depending on time, their function values $\beta_j(g_j(t_k))$ will first be normalized, and they will then be defined and fuzzified as *linguistic variable* β_j. The *linguistic values* "none", "low", "medium", "high", "extreme" strain and corresponding membership functions $\mu_\star(\beta_j)$ are assigned to them (figure 6 in section 4). The *modifiers* "decreasing", "constant", "increasing" and others can be used to subtilize the modelling. The same linguistic values apply for the fuzzified failure rate $\lambda_{(k)}$. Now m fuzzy sets $\mathbf{A}_j$ with

$$\mathbf{A}_j = \{\,(\beta_j\,;\,\mu(\beta_j)) \mid \beta_j \in [0;100\%] \subset \mathbf{R}\,;\,\mu(\beta_j) \in [0;1] \subset \mathbf{R}\,\} \tag{4}$$

are defined. The next step is the determination of rules, respectively implications, which are provided by means of experts' knowledge, e.g.: "IF the strain characteristic β_j is 'medium' AND the strain characteristic β_{j+1} is also 'medium' ist, THEN the failure rate λ is 'high'". Implications with two variables present a convenient speciality since they can be reasonably well represented in a matrix consisting of rules (figure 7 in section 4).

Based on the membership values $\mu_\star(\beta_{j,k}) > 0$ of the k-th time interval the *relevant* rules are chosen. These relevant rules define m subsets $\mathbf{A}_{j,\mathrm{sub}}$ on which the carthesian produkt $\mathbf{A}_{1,\mathrm{sub}} \otimes \mathbf{A}_{2,\mathrm{sub}} \otimes \ldots \otimes \mathbf{A}_{m,\mathrm{sub}}$ is applied to.

For determination of the membership function of an implication the MIN-method is used: The membership function of the THEN-part is limited to the temporary value of the IF-part. For the composition, namely the combination of all the relevant rules, a MAX-connection is chosen. The procedure of this MAX-MIN-inference is too conventional to be of any further interest here. An illustration of the procedure of an inference is presented in section 4. The inference provides a fuzzy quantity for the failure rate of the k-th time interval λ_k. The determination of a crisp value is done by means of the centroid defuzzification method. The abscissa of the centroid

$$\lambda_k = \frac{\int_a^b \lambda \cdot \mu(\lambda)\,\mathrm{d}\lambda}{\int_a^b \mu(\lambda)\,\mathrm{d}\lambda} \tag{5}$$

is the desired exact value for λ_k, with a and b as the upper and lower bound of the fuzzy quantity for $\mu_k(\lambda) > 0$. If λ_k is given, the constant failure rate function

$$h_k(t) = \lambda_k \tag{6}$$

can be determined.

For a more convenient notation the reliability characteristics are determined by means of the auxiliary function H_k with

$$H_k(t) = \lambda_k \cdot t + c_k \,. \tag{7}$$

It is the integral of h_k with the integration constant

$$c_k = \sum_{\kappa=1}^{k} \left((\lambda_{\kappa-1} - \lambda_\kappa) \cdot T_{\kappa-1} \right) \tag{8}$$

For the failure probability density function (pdf) f_k follows

$$f_k(t) = \lambda_k \cdot \exp\left[-\sum_{\kappa=1}^{k} H_k(t) \right] \quad \text{with} \quad T_{k-1} < t \le T_k \,. \tag{9}$$

For the corresponding failure cumulative distribution function (cdf) F_k is

$$F_k(t) = 1 - \exp\left[-\sum_{\kappa=1}^{k} H_k(t) \right] \quad \text{with} \quad T_{k-1} < t \le T_k \,. \tag{10}$$

APPLICATION: For illustrating the approach an elastic rotor with a transverse crack will be discussed (figure 4), simplified as a *laval-rotor*. The damage in form of a transverse crack of the rotor influences not only the behaviour of vibration but also the reliability of the system.

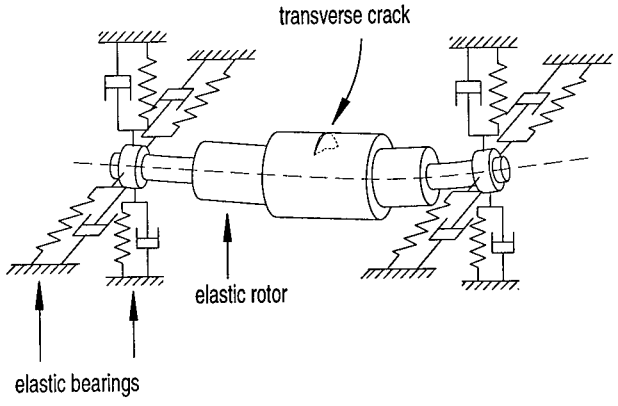


Figure 4 Example of an axle with a crack

A crack leads to a decrease in the cross-sectional area of the axis. This causes a change of the moment of resitance *w*. *w* enters inverse proportional into the first process characteristic which is the bending stress σ.

There is

$$\sigma(w) = \frac{d^2 \cdot \pi \cdot g \cdot \rho \cdot L^2}{16 \cdot w}. \tag{11}$$

The second process characteristic *torsional stress* τ with

$$\tau(p_{wr}) = \frac{p_{wr}}{2 \cdot \pi \cdot n \cdot w_p} \tag{12}$$

dependents on power (wattage) p_{wr} . It is: d - diameter, g - acceleration due to gravity, ρ - density; L - length of rotor, n - number of revolutions, w_p - polar moment of resitance. The normalized strain characteristics $\beta_1(\sigma(w))$ and $\beta_2(\tau(p_{wr}))$ (figure 5) represent a *subjective* influence of both process chracteristics on the reliability of the system. It has to be noticed that the turbine is stressed fare below its fatigue limit and that no empirically determined characteristics are available in this field.

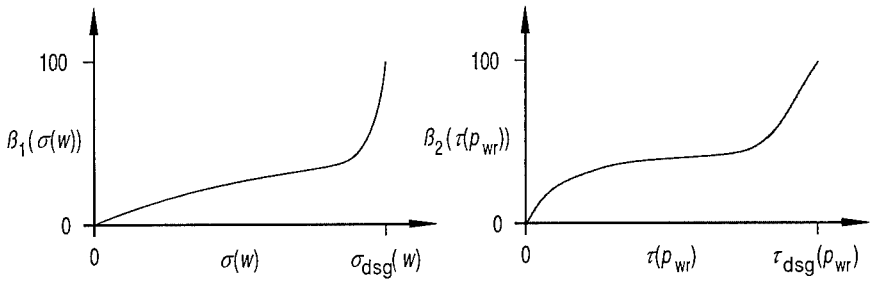


Figure 5 Formulation of strain characteristics $\beta_1(\sigma(w))$ and $\beta_2(\tau(p_{wr}))$. Design strains are denoted by $\sigma_{dsg}(w)$ and $\tau_{dsg}(p_{wr})$.

In the next step $\beta_1(\sigma(w))$ and $\beta_2(\tau(p_{wr}))$ are defined and fuzzified as linguistic variables β_1 and β_2. Linguistic values and the corresponding membership functions $\mu_\star(\beta_1)$, $\mu_\star(\beta_2)$ and $\mu_\star(\lambda)$ are assigned to β_1, β_2 and to the fuzzified failure rate λ (figure 6). Two fuzzy sets $\mathbf{A}_1$ and $\mathbf{A}_2$ are defined according to equation (4). The implications of a matrix consisting of rules are represented in figure 7. There emerge 25 rules for two variables with five values each.

If there is "no" strain, the failure rate in the example is still "small", to model the behaviour of aging. It can be seen in the asymmetric occupancy that the impact of β_2 on λ is slightly higher than the one of β_1. The classification was chosen monotonously increasing, which is, however, not necessarily so.

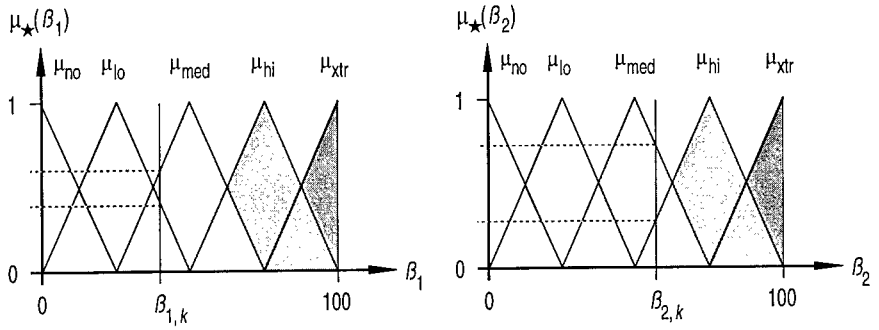


Figure 6 Fuzzified strain characteristics β_1 and β_2. λ is notated by analogy

β_2 ⟶

β_1 ↓	**none**	**low**	**medium**	**high**	**extreme**
none	low	low	medium	high	extreme
low	low	medium	medium	high	extreme
medium	low	medium	high	high	extreme
high	medium	medium	high	extreme	extreme
extreme	high	high	high	extreme	extreme

Figure 7 Martrix consisting of rules for the example discussed. The matrix represents the linguistic values of the failure rate. The hatched areas comprise the rules relevant for the given values.

In figure 6 some membership values have already been marked

$$\mu_{lo,k}(\beta_{1,k} \approx 40\%) \approx 0{,}4 \;;\; \mu_{med,k}(\beta_{1,k} \approx 40\%) \approx 0{,}6$$
$$\mu_{med,k}(\beta_{2,k} \approx 55\%) \approx 0{,}73 \;;\; \mu_{hi,k}(\beta_{2,k} \approx 55\%) \approx 0{,}27 .$$

They exemplarily represent the state of the k-th time interval. In this special case the four rules hatched in figure 7 are relevant:

❶ IF $\beta_{1,k}$ is "low" AND $\beta_{2,k}$ is "medium", THEN λ_k is "medium";
❷ IF $\beta_{1,k}$ is "low" AND $\beta_{2,k}$ is "high", THEN λ_k is "high";
❸ IF $\beta_{1,k}$ is "medium" AND $\beta_{2,k}$ is "medium", THEN λ_k is "high";
❹ IF $\beta_{1,k}$ is "medium" AND $\beta_{2,k}$ is "high", THEN λ_k is "high".

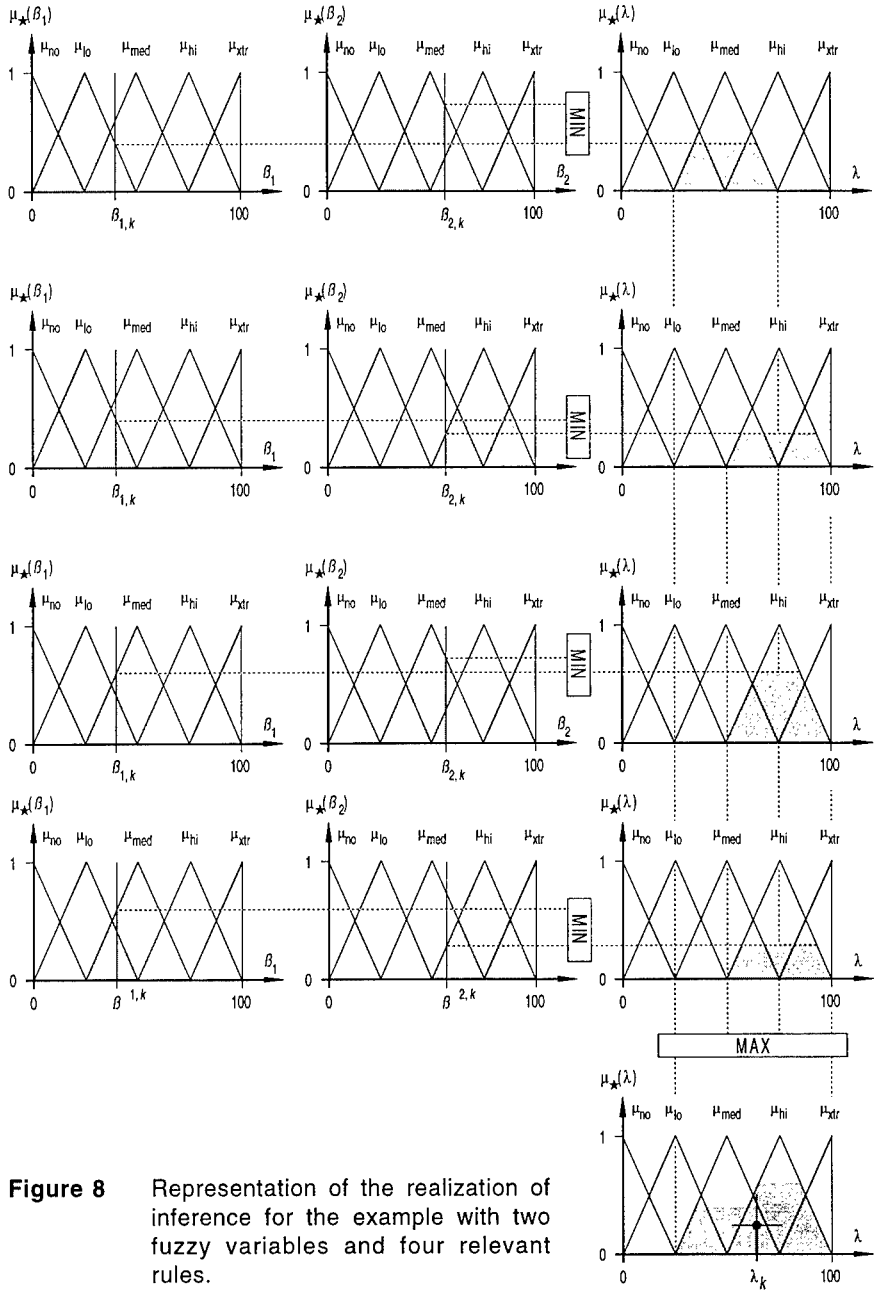


Figure 8 Representation of the realization of inference for the example with two fuzzy variables and four relevant rules.

These four rules define two subsets $\mathbf{A}_{1,sub}$ and $\mathbf{A}_{2,sub}$, on which the carthesian produkt $\mathbf{A}_{1,sub} \otimes \mathbf{A}_{2,sub}$ is applied to. The results are:

❶ $(\beta_{1,k} \approx 40\% ; \beta_{2,k} \approx 55\%) \rightarrow \min [\mu_{lo,k}(\beta_{1,k}) ; \mu_{med,k}(\beta_{2,k})] = 0{,}4$;
❷ $(\beta_{1,k} \approx 40\% ; \beta_{2,k} \approx 55\%) \rightarrow \min [\mu_{lo,k}(\beta_{1,k}) ; \mu_{hi,k}(\beta_{2,k})] = 0{,}27$;
❸ $(\beta_{1,k} \approx 40\% ; \beta_{2,k} \approx 55\%) \rightarrow \min [\mu_{med,k}(\beta_{1,k}) ; \mu_{med,k}(\beta_{2,k})] = 0{,}6$;
❹ $(\beta_{1,k} \approx 40\% ; \beta_{2,k} \approx 55\%) \rightarrow \min [\mu_{med,k}(\beta_{1,k}) ; \mu_{hi,k}(\beta_{2,k})] = 0{,}27$.

The next step is the application of the MAX-MIN-Inference. Figure 8 shows a detailed presentation of the forming of inference. The crisp value for λ_k determined with the centroid defuzzification method has already been entered. Avoiding complicated illustrations in figure 8, λ_k was also notated in the normalized way. However, an appropriate and common representation is the logarithmic presentation.
By application of the equations (6) to (10) the failure pdf and the failure cdf can then be computed. For this, figure 9 shows an exemplary course, but $h_k(t)$ does not refer to the example in figure 8.

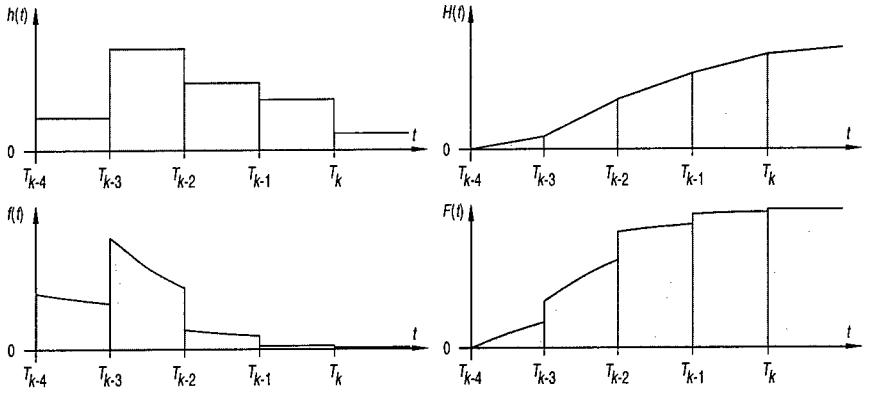


Figure 9 Example of a series of constant failure rate functions in equidistant intervals (on the left top) and the graphs of the corresponding functions *H*, *f* and *F*.

CONCLUSION: In this contribution the concept of conventional reliability analysis was left behind, in which the examinations of the system are completed before it is put into operation. The "*Safety and Reliability Control Engineering*"-concept (*SRCE*) is introduced which is continuously doing a so-called *real time reliability analysis*. This analysis makes it possible to determine data during the operation of the system, which might require a correction of the results from the analysis done before. The focus of this paper is the online-conversion of the signal data taken from the system into reliability characteristics. If the reliability

characteristics are provided, they are compared to the references. According to a control strategy the operating parameters will be evaluated. They then operate on the system to transfer it into the desired state.
The conversion is exemplary demonstrated at a mechanical system of an elastic rotor. First the process characteristics are transformed into subjectively defined strain characteristics. The dependence of the failure rate λ of the strain characteristic is formulated by application of fuzzy logic. Here a conventional MAX-MIN-inference is applied. It has to be mentioned that the quality of the modelling depends on the ability and the experience of the analysing team. This concerns most of all the establishment of the rule set, respectively the formulation of the implication and the estimation of the influence of the strain characteristics on the failure rate. If λ is given, the failure density and the failure probability can be easily determined.
By application of the *SRCE*-Konzepts an optimization of the maintenance strategies, a reduction of non-operation time, a resulting higher availability and, most of all, a higher safety of the system is reached.

REFERENCES:

[1] Bowles, J. E.; Peláez, C. E., "Application of Fuzzy Logic to Reliability Engineering", *Proceedings of the IEEE*, Vol. 83, No. 3, 1995.

[2] Dubois, D.; Parade, H., "Operations on fuzzy numbers", *Int. J. System Sciences*, Vol. 9, pp 613-626, 1978.

[3] "Fuzzy Sets and Applications", Selected Papers by L. A. Zadeh, R. R. Yager et al., Eds., New York, Wiley, 1987.

[4] Kersten, G., "Control and support of product and process design by methods of preventive quality management", *Steuerungen*, pp 19-23 ,September 1991, in German.

[5] Pouliezos, A. D.; Stavrakokis, G. S., "Real time fault monitoring of industrial processes", Kluwer Academic Publishers, 1994.

[6] Rückwald, R., "Modellgestützte Überwachung mechanischer Systeme mittels differentieller Parameteridentifikation", *VDI-Fortschrittberichte*, Reihe 8, Nr. 328, VDI-Verlag, Düsseldorf, Germany, 1993.

[7] Söffker, D., "A new model based tool for fault detection and isolation in machine- and rotordynamics", *Proc. 11th ASME Conf. Reliability, Stress Analysis, and Failure Prevention*, Boston, 1995.

[8] Söffker, D.; Rakowsky, U. K., "Perspectives of monitoring and control of vibrating structures by combining new methods of fault detection with new approaches of reliability engineering", submitted to: *12th ASME Conf. Reliability, Stress Analysis, and Failure Prevention*, Virginia Beach, accepted as abstract.

[9] Söffker, D.; Yu, T. J.; Müller, P. C., "State estimation of dynamical systems with nonlinearities by using proportional-integral observer", *Int. J. System Sci.*, Vol. 26, No. 9, pp 1571-1582, 1995.

[10] Zadeh, L. A., "Fuzzy Sets", *Information and Control*, Vol. 8, pp 338-353, 1965.

MONITORING RESIDUAL STRESS IN MACHINERY USING HOLOGRAPHIC TECHNIQUES

G. Salamo, G. Dovgalenko, Y. Onischenko, and A. Kniazkov
Department of Physics
University of Arkansas
Fayetteville, AR 72701

Abstract: We report the development of a nondestructive portable holographic technique for the detection of residual stress in materials. The technique we have used to achieve in-situ quantitative measurements is based on real-time holography. While the real-time holography concept has been demonstrated by other researchers, our approach is novel and results in a portable instrument which can carry out in-situ measurements. In all previous techniques for the measurement of internal stress the application of real-time holography had to be carried out on a large vibration isolation table and the results were only qualitative in nature. In the approach we will discuss we have used a technique which forces the reference and object wave and the recording material to move together when vibrators are present. In our experiment we have measured the internal stress in an aluminum structure using our holographic technique and find good agreement between our measurement, a measurement taken using a strain gauge, and computer simulations. The experiment was performed without vibration isolation using a HeNe laser, thermo-plastic holographic camera, and a small hole drilled in the aluminum structure. The two new features of our result are that the technique is portable and can be used in real situations and the technique provides quantitative, as opposed to qualitative, measurements of residual stress. As a result, the instrument can be used to monitor the development of residual stress in machinery.

Key Words: Displacement measurement, holographic interferometry; portable stress measurement, stress measurement

INTRODUCTION: One of the main difficulties in machinery evaluation is finding effective methods for evaluation. Methods are needed both during production and use. One method which has potential for meeting these needs is holographic interferometry.

Although holographic interferometry has been studied extensively for many years, it has not developed into an industrial tool for non-destructive testing. This is partially due to the technique's sensitivity to environmental disturbances, and partially due to the fact that interpretation of holographic data is too ambiguous.

Our effort, which we report on at this conference, is aimed at solving these problems. In particular, we have made significant progress on both environmental disturbances and on unambiguous data interpretation which makes possible portable detection for diverse applications as shown in Fig. 1.

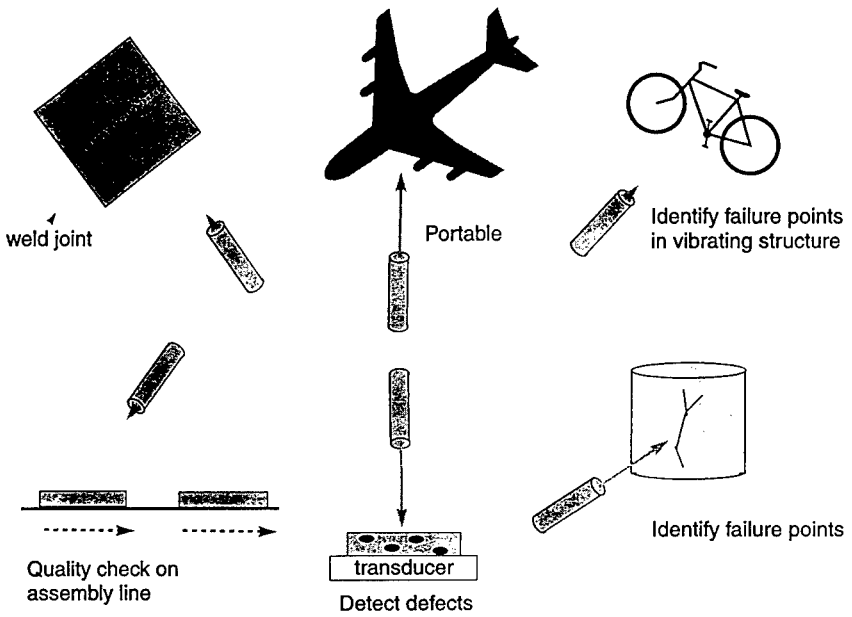


Fig. 1. Applications for holographic interferometry

HOLOGRAPHIC INTERFEROMETRY: The basic idea behind holographic interferometry is a measurement of displacement of a surface and an interpretation of that displacement measurement to deduce internal stress levels, fatigued areas, or defect locations. Since the technique is based on a measurement of displacement of a surface it is reasonable that environmental disturbances, such as vibrations, play a significant role in the effectiveness of the technique. Moreover, as a result of the fact that the displacement measurement must be interpreted in terms of stress, fatigue, or defects, it is reasonable that accurate interpretation plays a crucial role in the effectiveness of holographic interferometry as a testing device.

The holographic interferometry concept is simple to understand. Two light beams, one called the reference beam, the other the object beam, interfere on a holographic material as shown in Fig.2. This results in storage of the interference pattern in the holographic material. After storage, if the reference beam is redirected onto the storage material, both the object and reference beams, leaving the storage material are reproduced. If now the source of the object beam were to move slightly, a second interference pattern would be recorded. As a result, when the reference beam enters the storage medium two object beams are recreated and the interference between them results in an interference pattern which reveals the magnitude and direction of movement of the object.

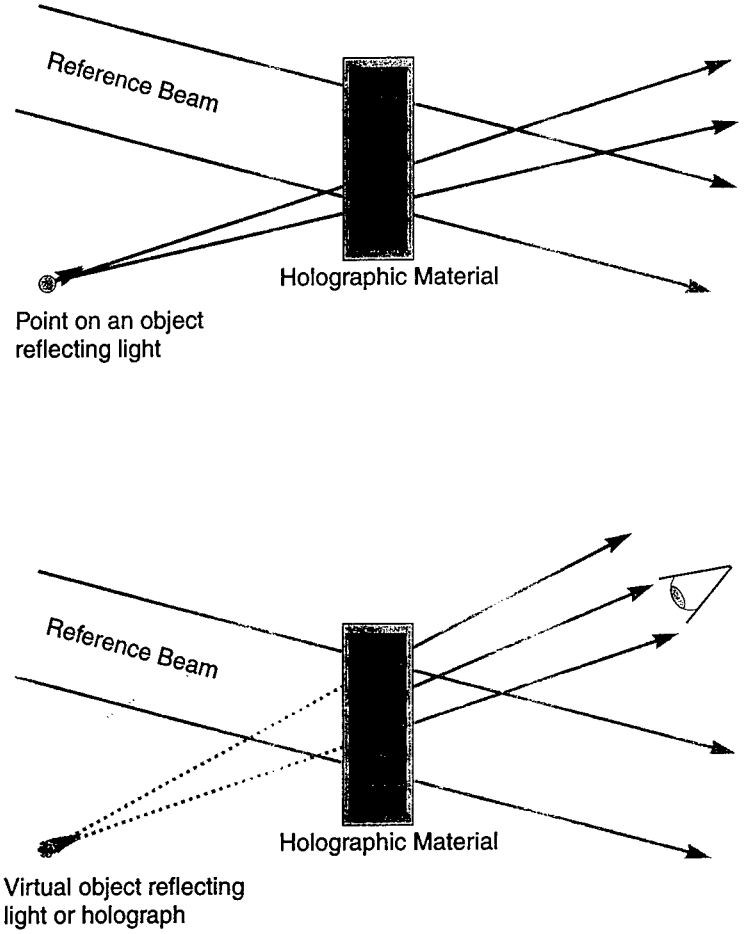


Fig. 2. Writing and reading a hologram

To use holographic interferometry to measure or monitor stress, fatigue or defects, a measurement of displacement or movement must be correlated to the degree and location of stress, fatigue or defect. Our approach to produce this correlation is based on scratching the surface of an object. A tiny, hairline scratch provides a trough into which neighboring material can move. Once this movement occurs, holographic interferometry can reveal the movement. The key point which correlates the measured movement to stress, fatigue, or defects is that if the material is perfect, then the removed material is unnoticed and no movement occurs.

In this paper we report on four measurements which indicate the progress and success of our research effort to develop an effective measurement device. In particular, we have (1) quantitatively measured stress in a thin film; (2) developed a measurement technique that can quantitatively image high frequency vibrations on a surface; (3) quantitatively measure stress using only a scratch on a surface as a load; (4) quantitatively sense failure points in tank like structures. We hope that the series of papers which we will present on these subjects will inspire possible collaborations between our team and scientists interested in developing techniques to monitor materials relating to machinery.

A typical holographic interferometry experimental set-up is shown in Fig.3. The reference and object beam from a tank form holograms in the thermoplastic holographic material. Reading out both holograms simultaneously then produces the interference pattern.

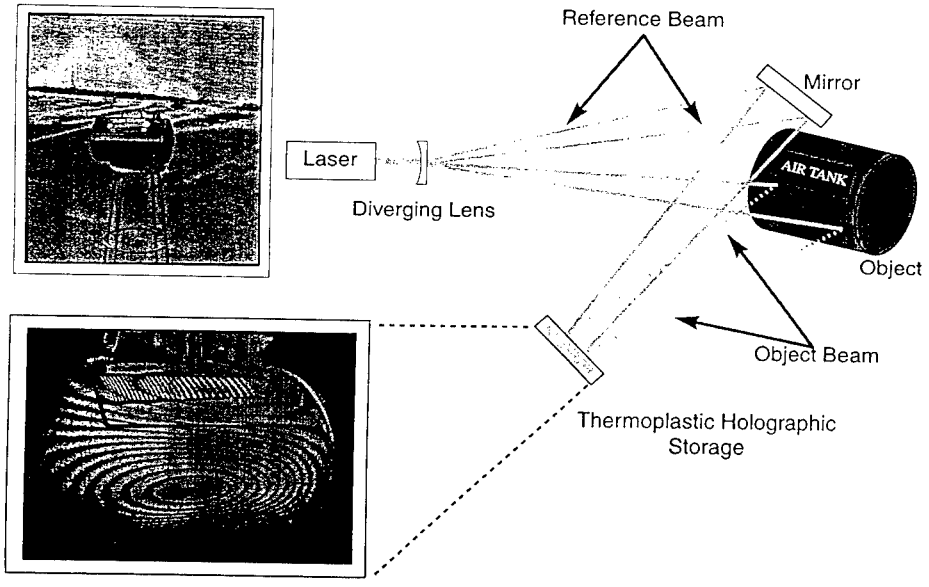


Fig. 3. Typical apparatus for measuring stress by holographic interferometry

MEASURING RESIDUAL STRESSES USING A SMALL BLIND-HOLE AND PHASE-SHIFT INTERFEROMETRY

Anatoli Kniazkov, George Dovgalenko, Gregory Salamo
University of Arkansas, 226 Physics Bld,
Fayetteville, AR, 72701

Tatiana Latishevskaia
Technical University, 29 Politechnicheskaia St.,
St.Petersburg, Russia, 195251.

Abstract: In this paper a new mathematical algorithm is proposed for non-destructive determination of residual stress by creating a small blind-hole with the depth of a scratch. Computer simulation of this algorithm has been made. A phase-shift technique in real-time reflection interferometry is used to measure the out-of-plane surface displacement surrounding a circular small depth blind-hole in a plate with the depth of a scratch. Comparison of the theoretical interference pattern with the experimental interferogram is presented.

Key Words: Algorithm; computer simulation; displacement measurement; interferometry; residual stress.

INTRODUCTION: The blind-hole-drilling method is a widely employed technique for non-destructive measurement of residual stress [1-3]. The procedure involves measuring surface deformations or displacements produced by the blind-hole in the plate where residual stress is present. Unfortunately, drilling a deep hole even if it is small in diameter is not exactly non-destructive. In this paper we present a new mathematical algorithm which is effective on a hole that is no deeper than a scratch on the surface.

Holographic interferometry has been most widely used as a tool for displacement analysis over many years [3-6]. One of the difficulties associated with the use of interferometric methods is the small phase change produced when using a small hole with the depth of a scratch. Another difficulty is the absence of accurate theoretical solutions for the radial and tangential components of the displacement for the case of a small hole with the depth of a scratch similar to the well-known Kirsch solutions for a deep hole.
We present a new theoretical solution for the components of displacement as a function of the hole deep, radius and thickness of the plate and a technique based on phase-shift interferometry to improve the sensitivity in order to determine the components of displacement in the case of a small blind-hole with the depth of a scratch on the surface.

MATHEMATIC ALGORITHM: We present a new solution to the problem of the displacement field associated with the introduction of a blind hole in a previously stressed material. In particular we considered three areas along the plate, as shown in Fig. 1. In the first area: (**I**) we have stressed the plate with a hole (the solution is well known), for the second area: (**II**) we have stressed the plate with a blind hole and we will find the solution, and in the third area: (**III**) we have stressed the plate without a hole (the solution is well known also). Referring to Fig. 1 and Fig. 2, the theoretical solution for the radial and tangential components of the average values of the stresses, $\widetilde{\sigma}_r$ and $\widetilde{\sigma}_\vartheta$, and for out-of-plane stress $\widetilde{\tau}_z$ respectively, near the circular hole produced in a infinite plate under biaxial state of stress (σ_1 , σ_2) are expressed in cylindrical coordinates as

$$\widetilde{\sigma}_r = \left(\frac{h+\Delta}{H}\right)\left[-\frac{a^2}{r^2}n - \delta\left(-\frac{4a^2}{r^2}+\frac{3a^4}{r^4}\right)\cos 2\theta\right]$$

$$\widetilde{\sigma}_\vartheta = \left(\frac{h+\Delta}{H}\right)\left[\frac{a^2}{r^2}n - \delta\frac{3a^4}{r^4}\cos 2\theta\right] \qquad (1)$$

$$\widetilde{\tau}_z = \left(\frac{h+\Delta}{H}\right)\left[\delta\left(-\frac{3a^4}{r^4}+\frac{2a^2}{r^2}\right)\sin 2\theta\right]$$

where $a = \frac{D}{2}$; $n = \frac{\sigma_1+\sigma_2}{2}$; $\delta = \frac{\sigma_2-\sigma_1}{2}$; $\Delta = \frac{\Delta_2-\Delta_1}{2}$.

The average displacements, respectively, near the circular hole are expressed in cylindrical coordinates as

$$\widetilde{U}_r = \frac{1}{E}\left(\frac{h+\Delta}{H}\right)\left[\frac{a^2}{r}n(1+\nu) - \delta\left(\frac{4a^2}{r} - \frac{a^4(1+\nu)}{r^3}\right)\cos 2\theta\right]$$

$$\widetilde{U}_\vartheta = \frac{1}{E}\left(\frac{h+\Delta}{H}\right)\delta\left[\frac{2a^2(1-\nu)}{r}+\frac{a^4(1+\nu)}{r^3}\right]\sin 2\theta \qquad (2)$$

$$\widetilde{U}_z = -\frac{4\nu h}{E}\left(\frac{h+\Delta}{H}\right)\delta\frac{a^2}{r^2}\cos 2\theta$$

where E - Young's modulus, ν - Poisson's ratio.

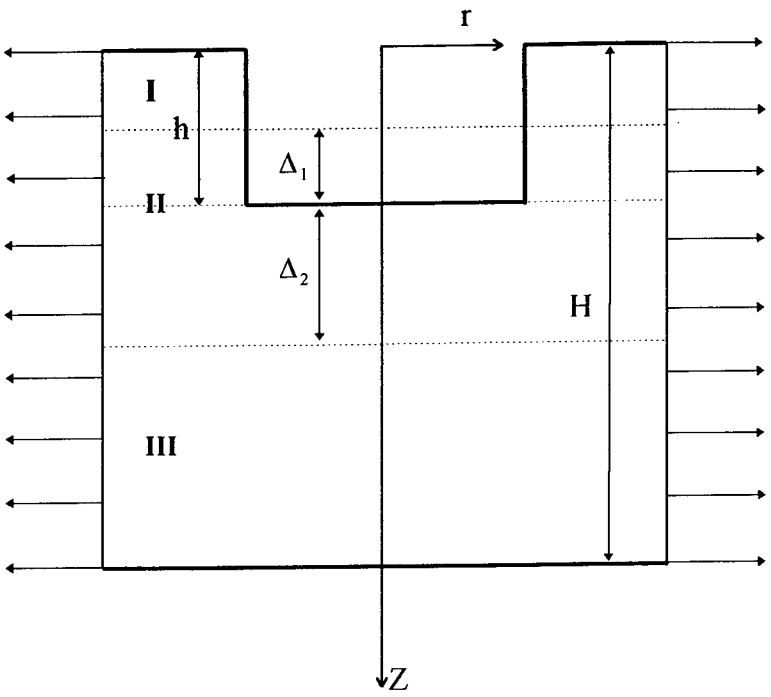


Fig. 1. The configuration of stressed the plate model as stressed the plate with hole (**I**), stressed the plate with blind-hole (**II**) and the stressed plate without hole (**III**).

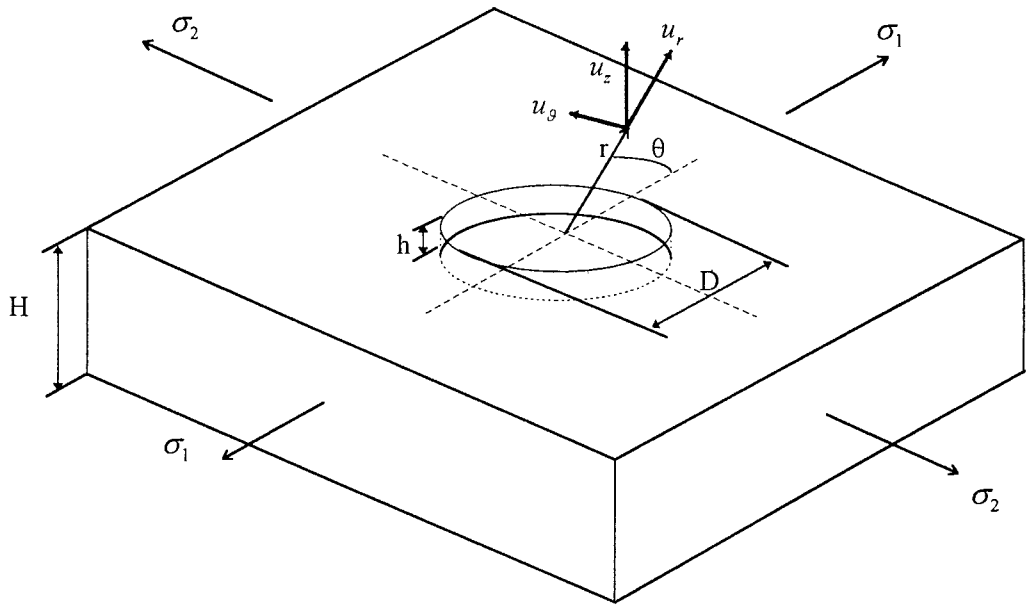


Fig. 2. Displacement components near the edge a blind-hole produced in a uniformly stressed plate.

To determine the principal stresses σ_1 and σ_2 we used the displacement equations on the border of a hole: $r = a = D/2$

$$\tilde{U}_r^a = \frac{D}{4E}\left(\frac{h+\Delta}{H}\right)[(\sigma_1+\sigma_2)(1+\nu)+(\sigma_1-\sigma_2)(3-\nu)\cos 2\theta]$$

$$\tilde{U}_\vartheta^a = \frac{D}{4E}\left(\frac{h+\Delta}{H}\right)(\sigma_2-\sigma_1)(3-\nu)\sin 2\theta \qquad (3)$$

$$\tilde{U}_z^a = \frac{2\nu h}{E}\left(\frac{h+\Delta}{H}\right)(\sigma_1-\sigma_2)\cos 2\theta$$

COHERENT OPTICS MEASUREMENT TECHNIQUES: We report the use of two coherent optics techniques. The first, holographic interferometry, provides the means to detect surface displacements, both in- and out-of-plane, produced by the introduction a blind hole in a material under stress. If a hologram is made as a double-exposure hologram, before and after blind-hole drilling, or viewed under real-time conditions after blind-hole drilling, the reconstructed wavefronts from the original image of the surface and from the deformed image of the surface after the introduction of the hole, will have a phase shift $\varphi(x, y)$ and some interference fringes will be observed by viewing this images through the hologram.

The interference pattern has been described the following equation:

$$A\tilde{U}_r + B\tilde{U}_\vartheta + C\tilde{U}_z = \varphi(x,y) = 2\pi k \qquad (4)$$

where k - interference number, A, B and C in the case of plane illumination wave referring to Fig. 3 will be:

$$A = \frac{2\pi}{\lambda}[\cos(\theta_0)\cos(\varphi_0)+\mathrm{sgn}(x)\cos(\theta_1)\cos(\varphi_1)]$$

$$B = \frac{2\pi}{\lambda}[\cos(\theta_0)\sin(\varphi_0)+\mathrm{sgn}(x)\cos(\theta_1)\sin(\varphi_1)] \qquad (5)$$

$$C = \frac{2\pi}{\lambda}[\cos(\theta_0)+\mathrm{sgn}(x)\cos(\theta_1)]$$

where λ - wavelength of the illumination source,

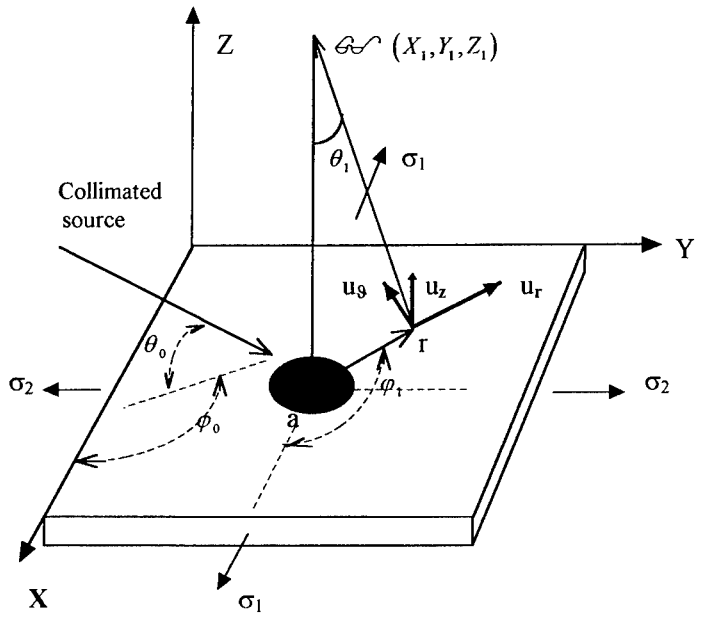


Fig. 3. Illumination and viewing coordinates for obtain shift interferogram

$$\theta_0 = ArcTan\frac{z_0}{\sqrt{(x-x_0)^2+(y-y_0)^2}} \qquad \varphi_0 = ArcTan\frac{y-y_0}{x-x_0}$$

$$\theta_1 = ArcTan\frac{z_1}{\sqrt{(x-x_1)^2+(y-y_1)^2}} \qquad \varphi_1 = ArcTan\frac{y-y_1}{x-x_1}$$

The second technique, digital phase-shifting interferometry, is direct phase shift $\varphi(x, y)$ measurement interferometry and can provide a fast and accurate means of obtaining surface displacement data. The application of this interferometry allowed investigations to be made of small blind holes with the depth of a scratch, when the phase shift $\varphi(x, y)$ is less than 2π. That is $\mathbf{k < 1}$ or when the interference pattern has less than one fringe.

The optical setup used in the tests is shown schematically in Fig. 4. Before drilling or scratching the surface, a hologram was made of a given location on the test object. After drilling (scratching) we obtain using a CCD-camera the resulting interference fringe pattern that is immediately viewed in real time on a video monitor of a PC-computer. Discrete phase shifts are introduced into the reference arm of the He-Ne laser after collimation using a piezoelectric transducer that is controlled by PC-computer. These interference patterns of the test surface area entered in the PC-computer where they are processed and the principal stresses of the test object are determined.

REFERENCES:

1. N.J. Rendler and I. Vigness, "Hole-drilling Strain-gage Method of Measuring Residual Stresses", *Experimental Mechanics*, Vol. **6** (12), pp. 577-586, 1966.
2. J.P. Sandifer and G.E. Bowie, "Residual Stress by Blind-hole Method with Off-center Hole", *Experimental Mechanics*, Vol. **18** (5), pp. 173-179, 1978.
3. D.V. Nelson and J.T. McCrickerd, "Residual-stress Determination Through Combined Use of Holographic Interferometry and Blind-hole Drilling", *Experimental Mechanics*, Vol. **26** (4), pp. 371-378, 1986.
4. A. Makino and D. Nelson, "Residual-stress Determination by Single-axis Holographic Interferometry and Hole Drilling - Part I: Theory", *Experimental Mechanics*, Vol. **34** (1), pp. 66-78, 1994.
5. S.-T. Lin, C.-T. Hsieh and C.-P. Hu, "Two Holographic Blind-hole Methods for Measuring Residual Stresses", *Experimental Mechanics*, Vol. **34** (4), pp. 141-147, 1994.
6. G. Dovgalenko, A. Kniazkov, Y. Onischenko and G. Salamo, "Fast Residual Stress Determination of Welded Joints using Reversible Holographic Interferometric - RHI Film", *Laser Interferometry VIII: Applications*, Proc. SPIE, Vol. 2861, pp. 32-40, 1996.

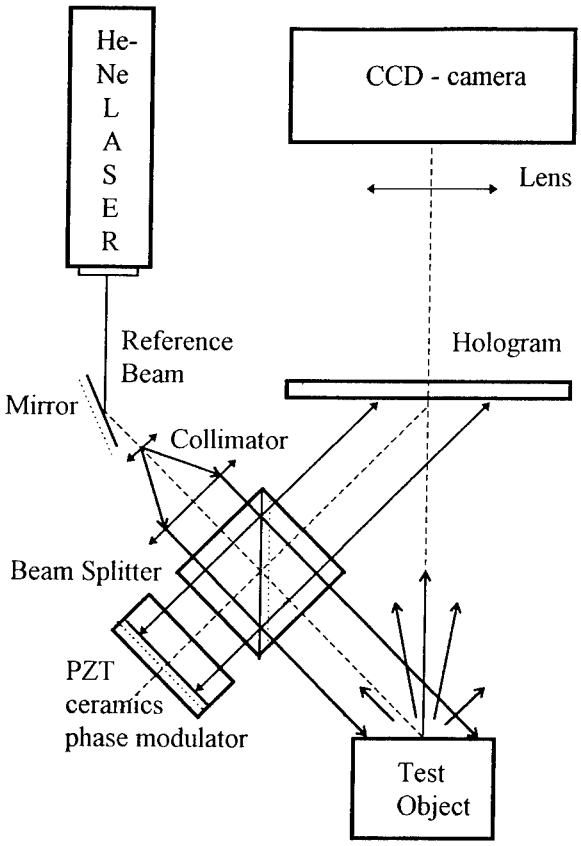


Fig. 4. Schematic of optical setup used in experiments.

TRANSITIONAL DATA FOR ESTIMATION OF GEARBOX REMAINING USEFUL LIFE

Carl S. Byington and James D. Kozlowski

The Pennsylvania State University
Applied Research Laboratory
State College, PA 16804

Abstract: Past failure tests conducted on mechanical equipment were directed towards the goal of determining the number of cycles to some measure of failure and/or a statistical measure of overall reliability such as a mean time to failure. True Condition-Based Maintenance (CBM), however, requires the identification and tracking of the sensor observables capable of detecting the faults and the ability to relate these variables to the overall health and remaining useful life of the machine. Thus, as part of a larger-scale research program in CBM for machinery, a Mechanical Diagnostics Test Bed (MDTB) was constructed to provide data on a commercial transmission as its health progresses from new to faulted and finally to failure. The paper presents a full description of the MDTB research station and its instrumentation. The generation of continuous run/good-to-bad gearbox transitional data will be used to determine appropriate data fusion and approximate reasoning techniques that result in the identification and detection of precursors to failure. The MDTB will also provide a test and evaluation vehicle for emerging prognostic time-to-failure/remaining life prediction algorithms and advanced sensors developed either at Penn State or by other participating researchers.

Key Words: Condition-Based Maintenance, Gearbox data, Mechanical failures, Prognostic techniques, Remaining useful life, Test design, Test monitoring, Transitional data

Introduction: The development of better maintenance practices is driven by the desire to reduce the risk of catastrophic failures, minimize maintenance costs, maximize system availability, and increase platform reliability. These goals are desirable from the application arenas of aircraft, ship and tanks to industrial manufacturing of all types. Moreover, given that maintenance is a key cost driver in military and commercial applications, it is an important area to focus our research efforts and improve the technology. Douglas [1] states that at nuclear power plants, the operations and maintenance portion of the direct operating costs (DOC) grew by more than 120% between 1981 and 1991 - a level more than twice as great as the fuel cost component. The EPRI M&DC [2] cites a *Forbes* magazine study that concluded that a third of the money spent on preventative maintenance in the utility industry was wasted. Correcting for the inflation rate, this translates to over $100 billion annually today.

The airline industry though has recognized the importance of incorporating predictive maintenance practices. [3] While time-based maintenance and statistical (reliability centered) measures are still the norm, condition assessment is increasing as the techniques allow. Built-in test equipment that provides self monitoring and diagnosis, non-destructive ultrasound, eddy current and x-ray have been used for some time to boost the ability of aircraft maintainers to detect emerging problems. As Feelar [4] describe, many mechanical failures can be eliminated with proper technology, procedures, and communication. This thread is common to almost every application domain.

CBM Techniques and Concepts: The primary research focus is the development of a prognostic capability, i.e., the ability to accurately and reliably predict the remaining useful life (RUL) of machinery in service. Hansen, et al [5] describe the limitations of past work that use probabilistic methods for failure prediction. Shortcomings associated with probabilistic distributions (Weibull, lognormal, etc.) include a lack of guidance towards the correct shape parameters and distribution for each failure mode. The result is often a hazard rate that may bear little or no resemblance to the actual circumstance. Application issues are compounded by the criticality of certain components.

Research and development has thus proceeded down two avenues. The first relates the appropriate microstructural material parameters to the macroscopic observables in physically-based way to a given failure. The second is to monitor the condition of the equipment on a periodic or semi-continuous basis and develop improved methods to analyze the precursors to failure. Such precursors will vary depending on the failure mode. To pry prognostications from such a broad range of scenarios, advanced decision-making methods will be necessary. One of the leading methods of combining individual sensors, data fusion, appears extremely promising for CBM development. Moreover, significant observable synergy is possible with digital intelligence techniques such as neural networks, fuzzy logic, and automated reasoning.

A roadmap of the technologies that will be required to achieve the objective of prognosis is shown in Figure 1, borrowed from Nickerson [6]. The top half of the process consists of monitoring the system in service while the bottom half details the analysis of data with a model-based approach. The monitoring includes both passive and active sensing systems. In active techniques, such as ultrasonic detection, a known signal is introduced into the structure and the response of the structure is monitored. Whereas with the passive approach, the self-generated vibration and noise are monitored. Accelerometers and acoustic emission transducers are examples of passive sensors.

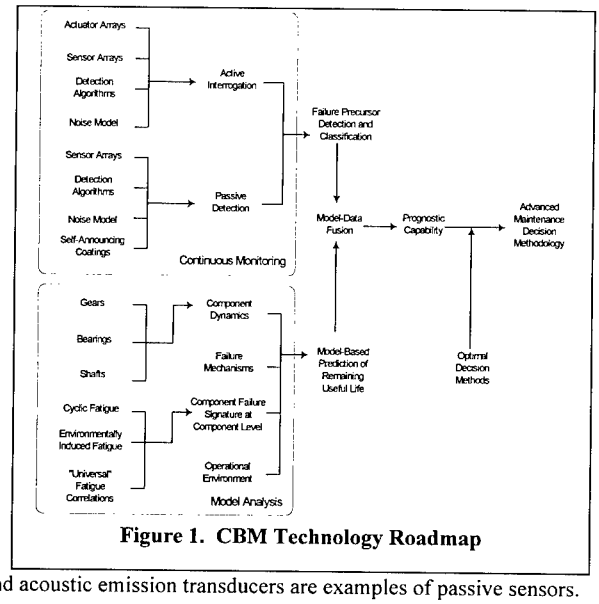


Figure 1. CBM Technology Roadmap

For mechanical systems, the dynamics of components such as shafts, gears, bearings, and housings are the primary concern. The mechanisms by which failure can initiate and how they evolve in these components can only be determined through an understanding of the signature that a particular fault generates and the way it translates into observables that can be detected. Finally, in order to accurately predict RUL, the operational environment in which the system will operate must be projected onto the system state.

An important hypothesis absorbed within the capability of predicting RUL is that failures in mechanical systems tend to follow a particular "failure trajectory" that may be predictable within a multi-dimensional state-space model sufficiently far in the future to be useful to the operator and to the maintainer. If this failure trajectory can be mapped and an account of the operational demand spectrum is proposed, then the ability to reliably and accurately predict the remaining useful life of equipment in service may be possible. Reference [7] provides greater detail.

Transitional Data Need: Advanced methods in maintenance technology are still in their infancy. Multiple new technologies have been applied to the maintenance and mechanical diagnostics problem. Examples include advanced detection methods for temperature, oil analysis, vibration signal processing, and many methods to monitor process parameters. A limiting factor in the further development of CBM has been and continues to be a lack of high fidelity data of faults as they initiate and evolve. This shortcoming is addressed by the MDTB effort in that it provides a realistic test stand that effectively represents an operational environment and is able to bridge the chasm between typical university-scale test facilities and the real world. It will be used to generate realistic data sets for use in development of CBM technology with the intention of making these data sets available.

Key to the current research project is that the ability to predict these failure trajectories for RUL estimation depends upon (1) the ability to accurately monitor a system via sensors, (2) the ability to process and fuse sensor data, and (3) the ability to model and predict the evolution of fault conditions. The MDTB directly accommodates the need for transitional data that tracks faults from initiation to an ultimate failure mode.

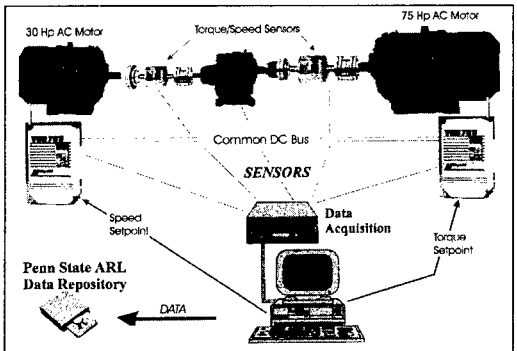


Figure 2. Mechanical Diagnostic Test Bed

Test Bed Description: The MDTB, shown in Figure 2, is functionally a motor-drivetrain-generator test stand. The gearbox is driven at a set input speed using a 30 Hp, 1750 rpm AC (drive) motor, and the torque is applied by a 75 Hp, 1750 rpm AC (absorption) motor. The maximum speed and torque are 3500 rpm and 225 ft-lbs respectively. The speed variation is accomplished by varying the frequency to the motor with a digital vector drive unit. The variation of the torque is accomplished by a similar vector unit capable of controlling the current output of the absorption motor. The system speed and torque set points are produced by analog input signals (0-10 VDC) supplied by the Data Acquisition (DAQ) Computer and a D/A board. The MDTB is highly efficient because the electrical power that is generated by the absorber is fed back to the driver motor. The mechanical and electrical losses are sustained by a small fraction of wall power.

The MDTB has the capability of testing single and double reduction industrial gearboxes with ratios from about 1.2:1 to 6:1. The gearboxes are nominally in the 5-20 Hp range. The system is sized to provide the maximum versatility to speed and torque settings. The motors provide about 2 to 5 times the rated torque of the selected gearboxes, and thus the system can provide good overload capability. The use of different reduction ratios and gearboxes than listed above is possible if appropriate consideration to system operation is given.

The motors and gearbox are hard-mounted and aligned on a bedplate. The bedplate is mounted using isolation feet to prevent vibration transmission to the floor. The shafts are connected with both flexible and rigid couplings. Torque limiting clutches are used on both sides of the gearbox to prevent the transmission of excessive torque as could occur with gear jam or bearing seizure. In addition, torque cells are used on both sides of the gearbox to directly monitor the efficiency and the loads transmitted. Figure 3 shows a picture of the hardware.

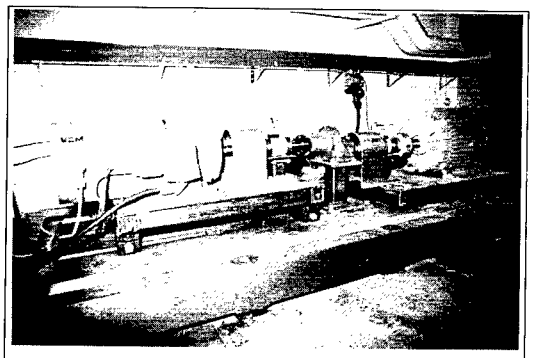


Figure 3. Photo of the Actual Test Bed

Instrumentation Design: Figure 4 illustrates the initially selected sensors and associated instrumentation. The figure also shows the different data acquisition cards and communication ports available on the PC.

Ten accelerometers and an acoustic microphone are placed on the test bed. Nine of these are single-axis, shear-type with a bandwidth of 20 kHz and the tenth accelerometer is a triaxial, shear-type with a bandwidth of 8 kHz. The triaxial accelerometer is included to determine whether triaxial data can provide significantly better sensor fusion for gearbox health assessment than the single-axis accelerometers. However, the measurement trade off is that the triaxial accelerometer possesses a lower frequency bandwidth than single-axis accelerometers. The microphone, placed in proximity to the test bed, provides a frequency range up to 22 kHz, which is almost twice the bandwidth of human audible range.

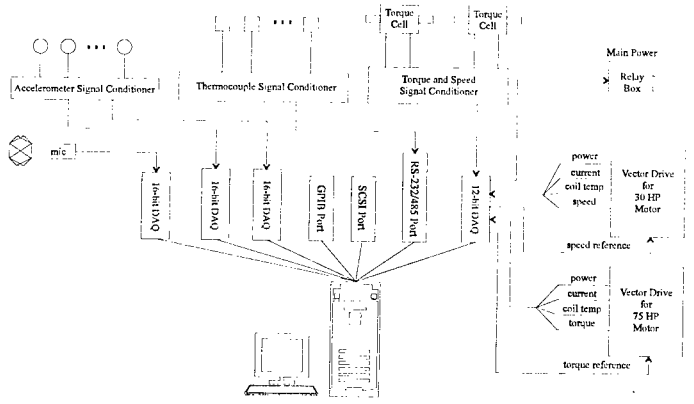


Figure 4. Instrumentation layout.

A total of 32 thermocouples are available for temperature readings on the MDTB. The thermocouples cover the gearbox and provide a low resolution image of surface temperature. A signal conditioner performs amplification, filtering, A/D conversion and transmits data to the PC.

Torque sensors are placed on the input and output sides of the gearbox. The torque cells also provide a speed measurement. Both torque and speed signals have a bandwidth of 500 Hz. The torque cell data can be used to measure the efficiency and power of the gearbox during the transition time before a failure occurs.

The vector drives, which control the drive and load motors, provide output signals that are sampled and stored on the PC. The output parameters include the: input power to the drives, root- mean-square (rms) currents, winding temperatures, motor speed and generator torque. As mentioned previously, the speed and torque set points for the vector drives are D/A signals from the PC. These signals allow automation and shutdown of the motors directly through the DAQ PC. As an additional emergency shutoff option, a relay box was placed in the main power line to disconnect the power from the vector drives and can be controlled manually or by the PC.

Sensor Level Structure: The choices for sensors used on the test bed and the priority associated with their introduction into the experimental program is given below. These assignments are based on practicality and effectiveness for analysis of fault prediction. The order in which the sensors are introduced on the test bed is divided into a five level structure:

- Level 1 sensors are required to operate and control the test bed. These sensors provide the means to set the operating conditions and determine if the station needs to be shut down.

- Level 2 sensors represent the most common and practical types of sensors used for diagnostics. Typical sensors in this level include those that measure vibration, temperature, and speed.

- Level 3 sensors represent a more expensive type of instrumentation and include torque sensors and audio microphones. Torque provides the transitional data for power and efficiency.

- Level 4 sensors are less common types of sensors such as acoustic emission (AE) transducers, which for instance, provide a measurement of the elastic stress waves generated at fatigue cracks. Given that this type of sensor requires a larger bandwidth with more expensive instrumentation, AE sensing is placed at this level. Infrared imaging is another example.

- Level 5 sensors represent sensors that are new or have not been used extensively for fault diagnostics. Sensors that detect different types of electromagnetic signatures or that use microscopic or localized imaging are examples of sensors at this level. Some Level 5 sensors may require modification of the gearbox to accommodate the sensors.

Data Acquisition and Storage: With the growing storage capacity of computer hardware, it was determined that the collected data should be stored in digital form. This choice allows the data to be more accessible for analysis as most of these tools are available as computer software.

Resolution and Sampling Speed: The conditioned signals of the MDTB sensors are all converted to digital data format with the highest resolution to which the sensors are accurate. To insure that the accuracy of the sensors is preserved, 16-bit data acquisition boards have been used for most of the measurements. When instrumentation was determined to have accuracy below 16 bits, a 12-bit data acquisition card is used. The sampling speed of the data acquisition cards is dictated by the frequency bandwidth of the instrumentation and sensors. The highest sampling speed required was 20 kilo Samples (kS)/s for the accelerometers and 44.1 kS/s for the microphone. The thermocouples are sampled 1 S/s by the thermocouple signal conditioner and then transmitted to the PC via a RS-232 serial port. The torque and speed signals and the vector drive monitoring signals only required a sampling of 1 kS/s.

Rolling Buffer and Event Triggers: Storage of continuously sampled sensors would require a massive amount of space and may only contain a small percentage of useful information. To

increase the percentage of useful information in the collected data and decrease the amount of required storage, a window of data is buffered in memory first. However, as the next set of samples are brought into the computer, the data in memory which represents the earliest set of samples are then deleted from the buffer. The buffered data is not stored immediately on the hard disk; it is first preprocessed to determine if it contains any new or noteworthy information about the system. One event that initiates storage of a block of buffered data is a periodic trigger. Accelerometer signals that exceed a threshold limit in the time or frequency domain also initiate storage of the data residing in the buffer. The preprocessed data for triggering is not saved; it is only used to determine if the buffered data should be saved.

Data Format and Storage: The data is stored in an unconditioned raw format with a header and index. The stored data is not converted to floating-point format. This choice minimizes the amount of storage required and floating-point conversion that would result in truncation errors. Header identify the general information about the data such as sampling rate, type and number of sensors, test condition, etc. An index points to the beginning of data blocks that could represent different sensor readings or time intervals.

The difficulty in using this format is that it may not be compatible with some commonly used databases. This format does however conform to the data standard being used within the overall CBM effort at Penn State. Converters are also being written so that the data can be imported into other desired databases. During each run of the test bed, the data is immediately stored on hard disk. Following the test run, the data will be moved to CD-ROM and/or digital tape.

Test Monitoring and Safety: Continuously run mechanical failure tests can require tens to thousands of hours to complete. To insure that the test bed will remain at full running capabilities, test monitoring and safety features are incorporated into the design of the test bed. Test monitoring allows the PC to determine if a problem arises during a test run that might damage the MDTB components or corrupt the data collection. Such monitoring allows the operator to be notified if a sensor has failed and requires replacement. Test monitoring also includes providing appropriate safety measures to prevent harm to personnel in the event of malfunction or gearbox failure during a test run.

Standard Methods: Some of the features within the scope of overall test monitoring are built into the vector control drives of the drive and load motors. Excessive speeds and/or loads will trigger the vector drives to stop the motors from driving the gearbox. The in-line clutches will decouple the motor from the gearbox if excessively high torques result on the input or output side of the gearbox. Dynamic braking is currently being added to the system to more quickly remove the stored energy in the motors and prevent damage to the rest of the test bed. A main power relay box provides a means of disconnecting all power to the MDTB manually or through the PC from the control room if the main breaker can not be reached quickly.

The PC and instrumentation are located in a control room thirty feet from the test bed. The test bed itself is located in a secured area to limit access to the machinery while the test bed is running. A guard shield is placed around the gearbox to prevent collateral damage in the event of a serious failure. Moreover, it is planned to add acoustic absorption panels around the MDTB to eliminate some noise and limit access from the work areas.

Smart Monitoring and Alarm: In addition to these standard measures, a smart monitoring procedure was designed to diagnose possible problems earlier and to quickly notify the operator. Early warning of a potential problem allows the problem to be corrected before it becomes irreversible. The early warning system allows the operator to leave the site. Figure 5 shows the smart monitoring procedure for the mechanical diagnostic test bed.

The procedure involves four computers connected to each other over a local area network (LAN) and/or by modems. The MDTB PC checks the test bed for problems and notifies the operator if any fault conditions or sensor problems are discovered. Depending on the severity of the problem, the MDTB PC will take different actions. An example of a low level alert would be a warning to the operator that disk storage is half full. The action for a low level alert would be for the PC to email the operator and log the event and its nature. If the operator is at the site, he can check the log directly. If he is not at the site he can check his email periodically from a different machine that has access to the mail server. An example of a mid-level alert is a loss of a thermocouple signal. The PC would email the operator over the network, page him using the modem, and log the event. A high-level alert or alarm would occur if there is a power failure at the test bed site. Since the PC is powered by an uninterruptible power supply (UPS), there is a brief period of time for the PC to email and page the operator, log the event, and shutdown. Email and pager notification is also performed when a test run has completed. If the PC itself crashes or locks up, the test bed could remain running unmonitored but will most likely coast down. If this situation does occur, it could result in the loss of data or damage to the test bed. In order to prevent this occurrence, another smart component was added to the system.

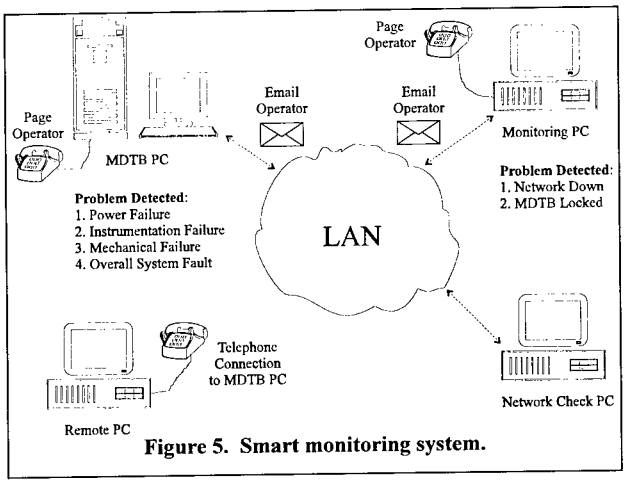


Figure 5. Smart monitoring system.

In addition to the MDTB PC controlling and monitoring the test bed, another computer connected to the MDTB PC monitors it via a LAN. This Monitoring PC attempts to communicate with the MDTB PC to determine if it is still active. A failure of communication indicates that either the network is down or the MDTB PC is not functioning. To check the network, a third computer, the Network PC, is incorporated. If communication to the third PC fails, a network failure warning is paged to the operator. At this point, the operator can wait for the Monitoring PC to verify the problem with the network as it attempts to connect to the MDTB via modem. The operator can also query the MDTB PC over the telephone lines using a Remote PC. If the Remote PC is unable to connect to the MDTB PC using a modem, the operator returns to the site to further investigate. In the event that the Monitoring PC successfully connects to the Network PC but fails to connect to the MDTB PC, indicating the possible failure of that computer, the operator is immediately paged. The operator may verify this notification by trying to connect the Remote PC to the MDTB PC. If connection fails, the operator returns to the site to check and possibly restart the station. The operational effectiveness of this smart monitoring procedure will dictate its use on future test beds.

Test and Analysis Plan: The test bed was constructed to collect data for both gear and bearing failures. Table lists some of the most common types of gear and bearing faults. In addition, to those listed, the gearbox may experience other types of faults. The main objective of

the test plan is to produce quantitative data that may contain failure precursors or warnings for a range of faults. Variations in operating parameters will cause different basins of fault conditions.

Gear Faults	Bearing Faults
Wear	Brinelling
Pitting	Spalling
Scoring	Retention failure
Fatigue crack	Cage failure

Table 1. Gear and bearing faults [8]

As stipulated in the MDTB description, the gearbox can be loaded at 2 to 5 times its maximum rating. By setting the torque at these levels will the same type of failure occur or can the type of failure be controlled by adjusting other operating conditions? The MDTB has two primary operating conditions that it controls: drive speed and load torque.

Given a low contamination level in the oil, the drive speed and load torque are the two major factors in gear failure. Different values of torque and speed will cause different types of wear and faults. Figure 6 illustrates the regions where failures are most likely to occur.

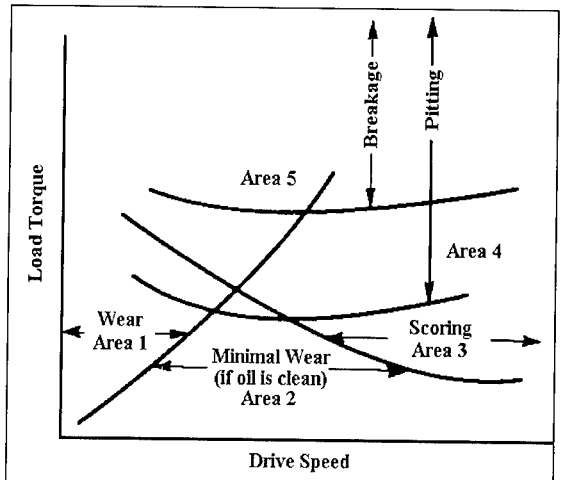


Figure 6. Regions of gear failures [9].

In Area 1, the gear is not running fast enough to develop an oil film, so wear occurs. In Area 2, the speed is fast enough and an oil film readily develops. The gears should be able to run with minimal wear. This area is most likely the best region to isolate bearing failures, which can be caused by imbalance load, inadequate lubrication or heating. In Area 3, scoring will take place since the load and speed are high enough to break down the existing oil film. Area 4 shows that with sufficiently high enough torque, pitting will occur. As the torque is increased even further, tooth breakage due to overload and stress fatigue will occur as shown in Area 5.

Based on the above discussion, the initial test plan will include test runs that should set the operating drive speed and load torque deep into each area to have transitional data for each of the possible faults. Primarily, Areas 4 and 5 will be the focus since they contain the faults which are more difficult to predict. In Area 2, an additional operating condition needs to be added to the test run to speed up the failure of the bearings without failing the gear. For example, a shaft imbalance causing high radial loads may result in bearing failure. Being able to control which type of failure occurs will allow control over the amount of data for each fault. If a particular type of fault requires more transitional data for analysis, adjustment of the operating conditions can increase the likelihood of producing the desired fault. While this is a fairly simple to

describe in such a theoretical fashion, it is expected to be one of the more challenging aspects of the experiments. This broad range is being pursued because the underlying goal of the experiments is to determine appropriate fault indicators and relate them to rate of damage for many failure states.

Figure 7 provides a graphical illustration of the current test plan. The two main variables are speed input and torque output. The torque percentages are relative to the specific nominal design torque. The matrix has been designed to collapse as appropriate. The first tests are those with fixed speed and torque using a 1.5 ratio gearbox. The goal is to progress to fully variable speed and torque ensembles.

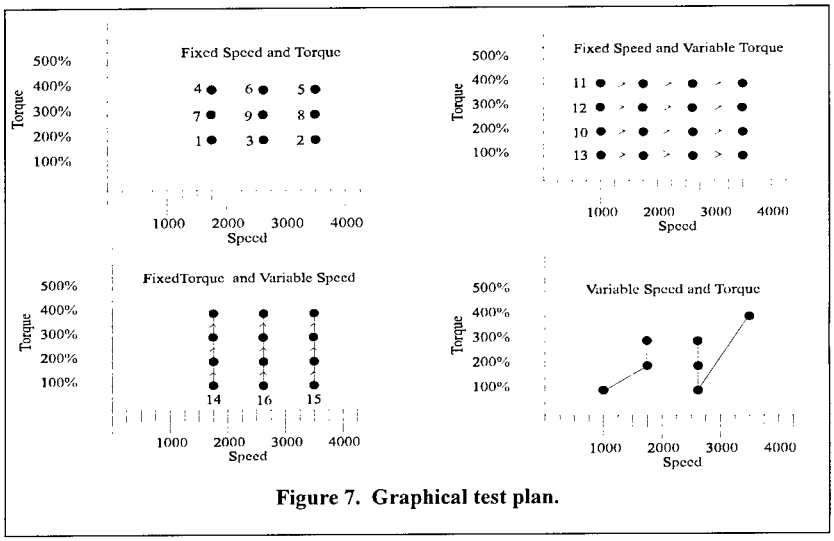


Figure 7. Graphical test plan.

Data Storage and Access: As mentioned earlier, data collected from a test run will be stored on CD-ROM or digital tape which will be available for analysis. In addition to the CD-ROM data library, a test run can be observed from the world wide web (WWW). The status of the test bed will be available as well as the most current sensor data from the run on the web page. The web page for the test bed allows researchers to gain a better feel for the test conditions and operation of the test bed. Such a page is currently available at Penn State at "http://wisdom.arl.psu.edu/".

Analysis Using Evolving MURI Toolkit: The primary purpose of collecting the data is to provide the analysis team with sets of various multiple sensor data over a range of test runs. The analysis will proceed along the classical lines [10], [11], [12] as well as the more advanced MURI techniques. As more data is collected, many of the analysis tools will be modified and improved based on this data. Also, the selection of sensors will be modified based on the results of the analysis. The test plan and the analysis tools are intended to evolve in a parallel fashion, building on the results in each area.

Summary: There is presently a lack of and need for transitional sensor data and multi-sensor data fusion that reflects the evolution of machine failures. The Mechanical Diagnostics Test Bed described herein was designed and built to provide a means to acquire such transitional data for

industrial transmissions. The test bed presently provides sensor data for vibration, temperature, torque, speed, and audio. However, based on previous test runs and analysis results, higher level type sensors such as acoustic emissions should be added to the test bed. Analysis plans include modeling these failure trajectories to be able to predict gear and bearing faults. The data and algorithms will be made available in an archive for future analysis and design of predictive diagnostics techniques. Collecting the data and developing algorithms in this manner will help to accelerate advances in CBM.

Acknowledgment: This research was supported by the Multidisciplinary University Research Initiative for Integrated Predictive Diagnostics sponsored by the Office of Naval Research (Contract Number N00014-95-1-0461). The authors are also grateful for support from the National Center for Advanced Drivetrain Technologies (NCADT) of the Applied Research Laboratory at The Pennsylvania State University. The Center is a nonprofit organization sponsored by the United States Navy Manufacturing Science and Technology (MS&T) Program under Contract Number N00039-92-C-0100.

References:

1. J. Douglas, "The Maintenance Revolution", *EPRI Journal*, pp. 6-15, May/June 1995.
2. EPRI M&DC, "Condition-Directed Maintenance", *Compressed Air Magazine*, pp. 34-39, August 1990.
3. "Aircraft Maintenance Cost Reduction", *Aerospace Engineering*, pp. 15-17, Jan/Feb 1996.
4. R. Feelar, H. Heule, R. Marlar, J. May, H. Sickels, "Aircraft Maintenance", *Industrial Maintenance and Plant Operations*, pp. 41-44, August 1996.
5. R. J. Hansen, D. L. Hall, S. K. Kurtz, "A New Approach to the Challenge of Machinery Prognostics", *ASME Gas Turbine and Aeroengine Congress*, June 13-16 1994.
6. G. W. Nickerson and D. L. Hall, "Research Imperatives for Condition-Based Maintenance", *Proceedings of the COMADEM Conference*, June 1995.
7. C. S. Byington & G. W. Nickerson, "Technology Issues for Condition-Based Maintenance", *7th Annual Predictive Maintenance Technology National Conference*, December 3-6 1995.
8. M. J. Neale, *Component Failures, Maintenance, and Repair*, Society of Automotive Engineers, Inc., Warrendale, PA, 1995.
9. D. P. Townsend, editor, *Dudley's Gear Handbook: The Design, Manufacture, and Application of Gears*, McGraw-Hill, New York, NY, 1992.
10. H. J. Rose, "Vibration Signature and Fatigue Crack Growth Analysis of a Gear Tooth Bending Fatigue Failure", *44th Meeting of the Mechanical Failure Prevention Group*, Virginia Beach, VA, April 1990.
11. R. M. Stewart, "Some Useful Data Analysis Techniques for Gearbox Diagnostics", *Application and Time Series Analysis Conference*, Southhampton University, Southhampton, England, September 1977.
12. J. J. Zakrajsek, "An Investigation of Gear Mesh Failure Prediction Techniques", *NASA TM 102340*, November 1989.

Navy Submarine Machinery Vibration Monitoring Cost/Benefit and Performance Analysis

G. Martin Milner
Principal Scientist
Analysis and Applied Research Division
Tracor Applied Sciences
6500 Tracor Ln.
Austin, TX 78725-2241

Russell R. Smith, Jr.
Code 7130
Naval Surface Warfare Center Carderock Division (NSWCCD)
Bethesda, MD 20084-5000

Abstract: Models are described for calculating curves of probability of detection (Pd) and probability of false alarm (Pfa) versus the threshold limits set for a vibration monitoring program (VMP). These values of Pd and Pfa are then used in a cost / effectiveness model based on the Bayes Criterion for binary hypothesis to examine the VMP cost benefit as a function of cost of unanticipated failure, cost for both necessary and unnecessary preventative maintenance or repair, and expected failure rates of the monitored machines.

Key Words: Machinery monitoring, cost benefit, sensitivity, specificity, failure rate

1.0 Introduction and Overview: The Submarine Monitoring, Maintenance and Support Program Office (SMMSO) administers programs to record vibration data periodically (approximately 100 day intervals) for most of the auxiliary rotating machinery on submarines. A primary goal of this Vibration Monitoring Program (VMP) is to reduce maintenance and repair cost by preventing catastrophic machine failures and through appropriate less expensive repairs, extend the machine service life. As budget constraints become more and more important, maintenance cost per year is considered one of the most critical evaluation factors.

The Naval Surface Warfare Center Carderock Division (NSWCCD) developed the current submarine VMP program for SMMSO and is responsible for data acquisition, analysis, and maintenance recommendations. The monitoring program designed by NSWCCD contains interdependent factors that combine to achieve the performance levels described here. The basic procedures are as follows.

1. Data are acquired at intervals of approximately 100 days on submarine machinery and vibration power spectra are analyzed to determine one of four conditions:
   a. Not flagged - The machine is operating properly and no action is required.
   b. Potential - The machine requires minor (i.e. low cost) maintenance followed by bi-weekly monitoring by ships force.
   c. Imminent - The machines require immediate major maintenance (i.e. high cost) most often including replacement of bearings.

d. Data Validity Problem - Data is not valid and must be reacquired.

2. Ships force then conducts the Imminent maintenance.
3. Ships force begins monitoring the Potential machines at bi-weekly intervals. The criteria (thresholds) and methods used in step 1 to determine Imminent problems are applied to all Potential machinery data at bi-weekly intervals. Machines that exceed Imminent levels during bi-weekly monitoring are repaired using the same Imminent recommendations used in step 1.

In general, the data are acquired at multiple machine positions and are processed in two frequency bands to obtain vibration spectra. These spectra are reviewed and assigned to the categories described in the above procedure.

Figure 1 illustrates important cost performance factors for the current NSWCCD VMP . The curves in figure 1 were derived from current estimates of VMP performance as measured by probability of correct detection (Pd) and Probability of unnecessary or false maintenance action (Pfa). Although unique VMP thresholds are defined for each machine type and are functions of frequency, these dependencies are normalized by measuring all vibration levels relative to associated maintenance thresholds. This normalization allows computation of the effect of an assumed constant shift in all machine thresholds as shown in figure 1.

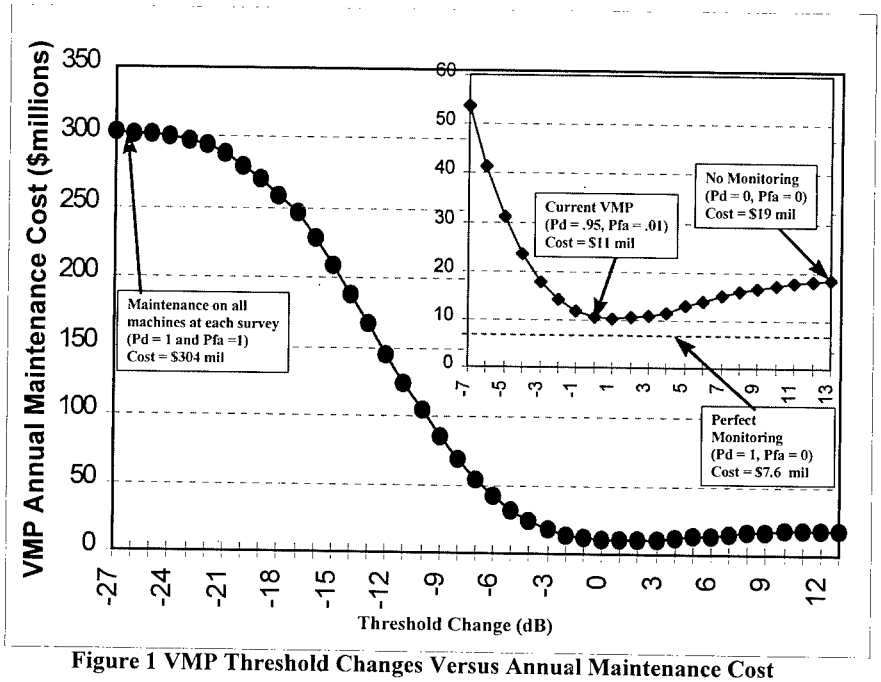


Figure 1 VMP Threshold Changes Versus Annual Maintenance Cost

Additional key program factors required to complete the calculation of the curve in figure 1 are: cost of a miss, cost of recommended maintenance (both necessary and unnecessary), total number of machines monitored, monitoring periodicity, and the expected submarine machinery failure rates. The estimated values for these factors are described later in this paper.

Important operational features illustrated by figure 1 include,

a. The expected annual cost if no monitoring is used and machines are run to failure. (No correct detection or false action or Pd = 0.0 and Pfa = 0.0)
b. The expected annual cost with a perfect monitoring system so that all calls for maintenance are at the correct time and no maintenance is performed sooner than required. With this hypothetical monitoring system the frequency of maintenance actions is equal to the failure rate of the machines, but the actions are less expensive than maintenance required if the machine fails (i.e Pfa = 0) and no machines are expected to fail without warning (i.e. Pd = 1.0).
c. The expected annual maintenance cost with the illogical system that calls all machines imminent failures and requires maintenance on all machines at every survey (Pd = 1.0 and Pfa =1.0).
d. The estimated annual cost of maintenance with the current VMP program which operates at values of Pd = .95 and Pfa = .01.
e. The expected annual maintenance cost as the vibration limit curves or thresholds shift by a constant value for all machines in the current VMP so that as threshold is decreased both Pd and Pfa increase (i.e more maintenance actions are performed) and as threshold is increased both Pd and Pfa decrease (i.e. fewer maintenance actions are performed).

Several important characteristics are evident from figure 1.

We are at first surprised by the sensitivity to threshold change in either direction. However, the effect is reasonable when we consider that the standard deviation of the data as measured with respect to these thresholds is about 5 dB and shifts on this order change the probabilities significantly. More is said regarding these effects in sections that follow in this paper.

Second, we see the dramatic difference in slope or sensitivity as threshold is decreased (Pd and Pfa go up) versus threshold increases (Pd and Pfa go down). This is explained by two factors; the failure rates of the machines are small so that relatively few machines are going to fail and need repair at any one survey (less than 3% of the population) and, the cost of unnecessary maintenance is not insignificant with respect to the failure cost. These two factors make the program much more sensitive to changes in Pfa as compared to changes in Pd. For the estimated miss and false action cost factors and failure rates, the annual maintenance cost is 26 times more sensitive to Pfa than to Pd. As an example, if for a new program, Pd increased from .95 to 1.0 (i.e. no misses), but Pfa also increased from .01 to .06, the annual maintenance cost would increase by over 14 million dollars per year. The change would more than double the annual maintenance cost.

Third, we see that the total savings possible from monitoring is approximately 19 million (at run to failure) minus 7.6 million (perfect monitoring) or 11.4 million. The current program operates at an annual maintenance cost of approximately 11 million; or approximately 8 million of the possible 11.4 million is saved in annual maintenance cost.

Finally, from the above, we conclude that program improvements must clearly show that Pfa or false actions are reduced as compared to the current program. Increases in detection probability can improve the program but the room for improvement is small with high risk. Cost models described in this paper were used to calculate that if Pd is increased from .95 to 1.0 (no misses) and somehow no increase in Pfa is observed, the expected annual maintenance cost would only decrease by 5.5% and would still be nearly 3 million above the minimum cost. Program changes that increase Pfa even slightly will get very expensive in a hurry and with only slight increases (from .01 to .035) the cost exceed the no monitoring cost.

Sections 2 through 6 of this paper describe the cost and performance models used to obtain the results described above and the methods used to estimate the required inputs. Sections 7 and 8 summarize the lessons learned regarding the analysis of VMP effectiveness.

2.0 Cost Analysis Description

The Bayes risk function for binary hypotheses (e.g. a good machine or bad machine hypothesis) is well known in detection and estimation theory [1]. The Bayes risk or, in this case cost equation, applicable to submarine VMP maintenance cost estimation is as follows,

Cost per survey = Pd x (Number Imminent failures per survey period) x (Cm)
+ (1-Pd) x (Number Imminent failures per survey period) x (Cfail)
+ (Pfa) x (Number good machines analyzed per survey period) x (Cm) (1)

where,

1. Pd = Probability that a bad machine (i.e. destined to fail in the next survey interval) is correctly detected and maintenance is recommended.
2. Pfa = Probability that a good machine (i.e. will not fail in the next survey interval) is falsely classified as needing maintenance.
3. Survey intervals are 100 days so that the average number of surveys conducted per year per submarine = 3.65 (The effects of survey interval are addressed in section 3.0)
4. Number of submarines in VMP = 95.
5. Average cost of repair when a machine fails without warning, Cfail = $25,000.00 per failure (See section 6.0).
6. Average cost of repair when the maintenance action recommended by VMP is executed C m = $10,000.00 per recommendation (See section 6.0).
7. Number of machines analyzed per survey = 100 machines per survey.
8. If 100 machines are monitored per submarine, it is estimated that maintenance is required on 8 of these machines each year. (Section 4.0 contains survey data and rationale supporting the value of 8 failures per year per submarine.)

The first term in Eq. (1) is the expected cost for bad machines that are detected prior to failure. The second term in Eq. (2) is cost of repair of machines that are missed (Note, Probability of a miss = 1 - Pd). The third term in Eq. (1) is the cost of false alarms (maintenance performed unnecessarily).

As baselines consider two cases. First, for no monitoring program (Pd = 0.0 and Pfa = 0.0). The cost equation then reduces to,

Cost per survey period = Number of Imminent failures per survey period x Cfail
= 208 x $25,000.00 = $5,205,478.00 per survey period, or
Cost per year = 3.65 x $5,205,478.00 = $19,000,000.00.
(Without monitoring and run to failure maintenance program)

Second, for a perfect monitoring system (Pd = 1.0 and Pfa = 0.0) the cost becomes,

Cost per survey period = 208 x (Cm) = 208 x $10,000.00
= $2,082,191.00 per survey period or
Cost per year = 3.65 x $2,082,191.00 = $7,600,000.00 per year.
(For a perfect monitoring program)

From this illustration we conclude that for the assumed values , a perfect monitoring program would save the Navy approximately $11,400,000.00 per year in maintenance cost.

The SMMS0 VMP thresholds have been adjusted so that approximate values of Pd and Pfa are: Pd = .95 and Pfa = .01 (Section 5.0 describes examples of prior studies that estimate Pd and Pfa). Using these values in the cost equation yields,

Cost per survey period = .95 x 208 x $10,000
+ .05 x 208 x $25,000.00
+ .01 x (97.8 x 95) x $10,000.00
= 1,978,081.00 + 260,274.00 + 929,178.00
= $3,167,533 per survey period, or

Cost per year = 3.65 x 3,167,533.00 = $ 11,561,500.00 per year
(For Current SMMSO VMP program)

This illustration indicates a net maintenance cost savings of $7,438,500.00 per year for current VMP performance levels as compared to a run to failure program.

The calculations presented above do not account for the survey recommendations for minor preventative maintenance (such as check lubrication). These "Potential Problem" recommendations serve two very important purposes. First, they are a prescreen for on-ship monitoring and dramatically reduce total program false alarm rate. Second, the initial maintenance often extends the time between Imminent failures significantly. The

first effect has been incorporated in the false alarm calculations described in this paper. However, the effects of "Potential Problem" maintenance on extended time between Imminent failures is a significant cost savings that should be added to all cost savings calculations presented in this paper.

3.0 Effect of Survey Interval and Cost versus Threshold

Recent calculations have indicated that on average, 3.2 surveys per year are performed on each submarine. A first approximation for the recent statistics is that the total expected number of Imminent failures per submarine per year remains the same so an average of 2.5 (calculated by 8 / 3.2). Imminent failures per submarine per survey are expected. The detection and false alarm probabilities are also assumed to remain the same (Note, the values of Pd and Pfa used in this analysis were based on a 100 day monitoring interval or 3.65 surveys per year per submarine. Analysis has shown that Pd and Pfa change less than 5% for 3.2 surveys per year per submarine versus 3.65 surveys per year per submarine.)

4.0 Failure Rate Estimate

Several methods were used to estimate the number of imminent failures expected per submarine per survey. Three of these are discussed here. First we know the number of total recommended maintenance actions per survey and independently the probabilities of detection and false alarm (see section 5); from these we would estimate 2 Imminent failures per submarine per quarter. Second the bearing set replacement rates have been tracked for submarines and are typically .3 bearing sets per submarine per month which gives a value of approximately .9 Imminent failures per submarine per quarter; however, the reporting is known to be about half which yields a number of 1.8 Imminent failures per quarter. Third, as a part of VMP threshold studies performed by NSWCCD from 1989 through 1991, NSWCCD surveyed 21 submarines over a period of one year. A total of 11 failures of machines (included in the SMMSO program) were observed and counted as missed by the program. This converts to .52 missed failures per submarine per year. If we use a value of .05 for probability of a miss (Pmiss =1- .95) and a value of 8 failures per year per submarine, the expected miss rate is .05 x 8 = .40 misses per year per submarine. The prior NSWCCD study set the 90% confidence bounds on miss rate between .32 and .89 missed failures per submarine per year. Therefore, the observed missed failure rate is very near the theoretical value for Pd = .95 and the assumed failure rate of 8 per submarine per year.

5.0 Values of Pd and Pfa

Pd

The probability of detection or Pd has been estimated in a number of studies and by a variety of methods in prior studies by NSWCCD.

In one previous study, the trim and drain pumps were of primary concern, since Navy personnel suggested that the thresholds should be raised significantly to reduce what was perceived to be unnecessary maintenance. Data were collected and processed to obtain statistical descriptions (i.e. mean and standard deviation) of the levels versus time as

measured with respect to the "Imminent thresholds" for machines known to fail on specific dates. These time varying descriptions were used as the basis for calculating the total program cumulative probabilities that the levels would exceed "Imminent thresholds" with single samples for laboratory analysis and bi-weekly samples intervals by ships force. The single sample probabilities of assignment of the machines to bi-weekly sampling or "Potential " calls were also considered. The combined effects of all major VMP procedures were accounted for in the models. These models allowed NSWCCD to determine the program sensitivity to threshold adjustments. The results showed a surprising sensitivity to threshold. Adjustments of less than 5 dB can cause Pd to drop from a range of .85 to .95 to a range from .3 to .65. The estimates of Pd ranged from .85 to .98 depending on the data analyzed.

A second, more direct method, of estimating Pd was to calculate the rate of undetected catastrophic failures with and without VMP. If Pd is on the order of .9 for the VMP, we should find that the undetected failure rates for machines monitored by the VMP is approximately 10 times less than machines that are not monitored by the VMP. Two prior studies examined failure rates of machines with and without VMP. A ratio of failures rates of 6.15 was estimated with a 90 confidence interval between 3 and 14. This corresponds to Pd = .84 with limits between .67 and .93. In a second identical survey a value of 9.2 for the ratio of failure rates was measured and converts to a value of approximately Pd = .90.

Both methods of estimating Pd discussed above may yield values that are biased lower than the true values. Both estimates use primarily data corresponding to machine failures only. It is reasonable to assume that data obtained for machine failures may have levels that are more difficult to detect than data for a pool of imminent machine failures. By definition the monitoring program failed to prevent the machine failures and some, if not most, of these may have been a result of unusually low vibration levels. For studies that examine possible program changes and effects, we believe that the larger values obtained in these studies are more appropriate. Therefore, we selected Pd = .95 to illustrate the effect of threshold variations.

Pfa

The first methods described above for estimating Pd were also used to estimate the value of Pfa. The data for trim and drain pumps suggested a value of approximately Pfa = .05 for the total false alarm probability (including laboratory analysis and bi-weekly monitoring by ships force). An analysis of a relatively small sample of fans gave a value of Pfa = .02 for the data collected between 1985 and 1989. Subsequent to the 1990 study, the NSWCCD analysts have significantly reduced Pfa as a part of the effort to deliver automated systems to submarine base sites. The thresholds have been adjusted at selected frequency regions to reduce Pfa. Throughout this process, the data for all submarines for 3 to 4 years has been completely reprocessed (over 30,000 measurements) to insure that no prior imminent problems were missed. Thus, this process has reduced Pfa significantly and Pd has remained constant.

A second independent method was also used to estimate Pfa. If we assume a known value of Pd and a known value of failure rate we can calculate the value of Pfa if we know the average number of machines called Imminent failures per survey. We have a very good estimate of the number of Imminent machines through review of prior messages. For all SSN data collected from 1992 through 1995 NSWCCD analyst Imminent machine call rate ranged from .8 per survey to 2.0 per survey depending on the specific site. The average over all sites was 1.55 Imminent machine calls per survey. A recent survey of the maintenance feedback obtained as a part of the VMP was completed by NSWCCD to determine the average number of Imminent calls per submarine per survey interval made by ships force as a result of bi-weekly monitoring. The average number of repairs or Imminent calls by ships force (including all classes of submarines) was 1.67 per survey interval. Thus, an average total of 3.22 (1.55 + 1.67 = 3.22) Imminent calls per submarine per survey interval are made by the NSWCCD VMP process. Subtracting the number of expected failures per survey interval detected by the VMP yields 3.22 - (.95 x 2.5) = .845 additional repairs per submarine per survey interval. If this number is divided by the expected number of good machines which is 100 - 2.5 = 97.5, the estimated value of Pfa is, Pfa = .845 / 97.5 = .0087. Very near the value of .01 used in this analysis. A value of Pfa = .01 is required to be consistent with other known parameters such as imminent machine message rate and machine failure rate. We therefore use Pfa = .01 to illustrate VMP performance effects.

6.0 Cost of a Miss and Cost of a False Alarm

NSWCCD has attempted to obtain estimates for an average of all machines from personnel at repair facilities and other sources. In this survey NSWCCD has asked that secondary cost such as connected failures or delays be included as well as the direct labor and materials costs. The values of $25,000 for a miss and $10,000 for a NSWCCD imminent machine call have been presented to all SSN submarine base site personnel. To date, NSWCCD has received general agreement that the values are reasonable.

7.0 Discussion of NSWCCD VMP Program And It's Effectiveness

We estimate that for each dollar spent on the current VMP, approximately 8 dollars are saved in annual maintenance cost. In total, for a program implementation cost of less than 1 million the Navy saves 8 million per year. These savings are a direct consequence of the very high performance levels of the monitoring.

The medical profession is highly dependent on a variety of techniques or test procedures to monitor and detect diseases and the two key performance measures are Sensitivity and Specificity. Sensitivity is exactly equal to the term Pd as described in this paper and Specificity is equal to 1 - Pfa as described here. Often medical detection and diagnosis must be applied to populations with very few diseased patients but with cost of both miss and false alarm very high. These situations require very high Sensitivity (high Pd) and high Specificity low Pfa) to be of value. Recently the National Institute of Health (NIH) requested a presentation of the methods used in the current VMP and their applicability to monitoring of permanently implanted heart pumps. The program initiated in 1995 by NIH to develop such an application will clearly need monitoring with exceptionally high

Specificity and Sensitivity. The values reported in this paper of .95 for Specificity and .99 for Sensitivity are thought to be at the levels required.

As described previously in this paper (see section 1), the monitoring program designed by NSWCCD is deceptively simple but contains interdependent factors that combine to achieve the performance levels described here.

Imminent levels are set to detect severe faults in the last few months of operation. These levels are designed to keep the incidence of unnecessary action very low. We estimate that only approximately 40% of imminent failures are detected during step 1 or the initial laboratory survey analysis. The Potential category is used to select a small population of machines that are very likely to include the remaining 60% of the machines expected to fail prior to the next survey. Approximately 10% of the machines are placed in the Potential category. If ships force monitored all machines using the Imminent criteria and surveyed all machines at bi-weekly intervals we would expect the false actions resulting from random variations in data to be many times higher than the single survey results since the data is surveyed approximately 8 times between surveys. The Potential category acts as a screen or filter that limits the number of machines monitored bi-weekly. This reduces the bi-weekly false action rate by approximately a factor of ten. Since the cost of false assignment is relatively low, the Potential criteria is designed to insure detection of possible future failures at the expense of a large number of false assignments to Potential. We estimate an average of only 1.8 Potential machines per submarine per survey become Imminent during the subsequent bi-weekly monitoring but that the probability that the true Imminent failures are contained in the Potential list is very near 1.0.

Since the ships force applies the Imminent criteria to all Potential machines, a strong interdependence between these criteria and final program performance exists. Figure 2 illustrates these interdependent effects. The curves in figure 2 are based on an assumed value for single survey detection of Imminent machines of .4 (the approximate value for the current VMP) and the cost and machine failure parameters described in sections 1 and 2. In figure 2, the values for Pfa are the combined values for both laboratory analysis at periodic survey intervals and ship analysis at bi-weekly intervals. As the assumed laboratory criteria changes and the number of Imminent machines called (Nimm) changes, the ships force frequency of Imminent calls goes up proportionately since we assume the same criteria is applied bi-weekly. As the number of Potential calls goes up, this criteria is applied to more machines at bi-weekly intervals. Thus, there is a compounding of effects. The current operating point for the NSWCCD VMP is near the lower curve (Nimm = 1.5) and at approximately 10 Potential calls per survey.

8.0 Evaluation of Alternative Monitoring Programs

Although we have shown room for improvement exists for the NSWCCD VMP, the risk of degradation is high and slight changes in performance can result in loss of millions of dollars. The improvements are much more likely to come from decreases in Pfa rather than increases in Pd.

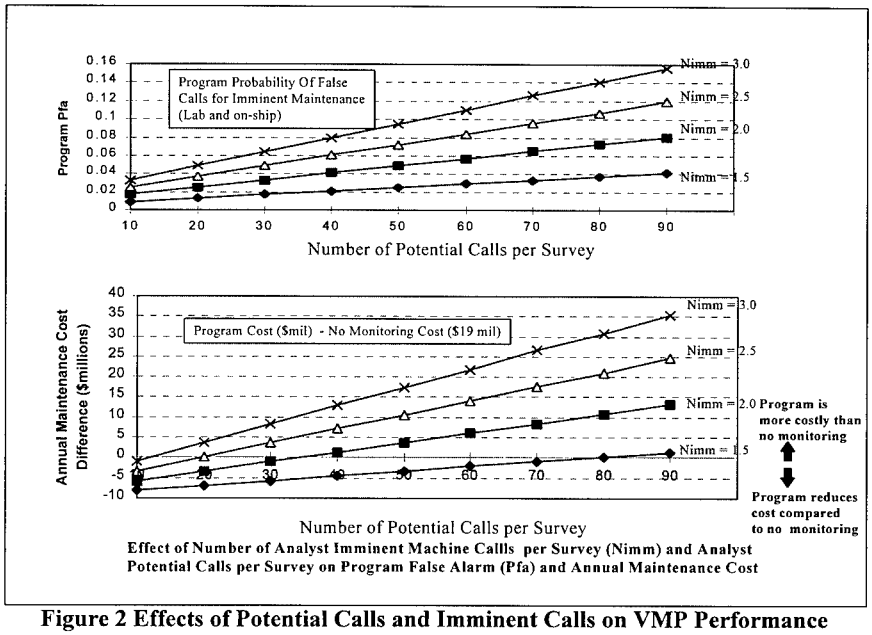


Figure 2 Effects of Potential Calls and Imminent Calls on VMP Performance

In a prior study funded jointly by NASA and SMMSO [2] experiments were conducted to collect an extensive database of time domain vibration data for a variety of machinery problem types and for a statistically significant number of repeated trials. A goal of this study was to investigate the relative performance of a variety of alternatives to vibration spectrum analysis as currently performed in the NSWCCD VMP program. Processing investigated included: standard baseband spectral analysis, high frequency demodulation, general time-frequency analysis, cepstrum, process modeling techniques, Kurtosis, rotation synchronized processing to achieve once per revolution averages, likelihood ratio processing of time domain data as well as frequency domain data, and others. The data collected and the analysis results of this study allowed us to evaluate techniques by either fixing Pd and selecting the minimum Pfa or fixing Pfa and selecting the maximum Pd. The number of experiments was sufficient to quantify the experimental errors in the estimates of these quantities and to insure that meaningful performance differences were measured. The study demonstrated that although for specific problem types other techniques may achieve higher Pd or Sensitivity than spectral analysis, spectral analysis provided the highest Pd and lowest Pfa as averaged over problem types and allowed an assessment of level of severity.

A key factor missing from the result described above was that the faults could not be directly related to ultimate failures. The key question was: When should a fault be repaired? Maintenance monitoring must address this question. The issue is NOT: Can we

detect a slight change in operating conditions which all machines show within hours of operation? Again, the question is: can we say precisely when the machine should be repaired so that high reliability is achieved and our cost is minimized?

The NSWCCD monitoring program is able to answer failure question primarily because it has operated for over 25 years with constant feedback regarding calls that were made unnecessarily and machines that failed without warning. Both types of feedback resulted in possible threshold changes or procedural changes to correct the problem. It is our belief that a monitoring program can not be developed from theory alone. It must be developed from the machine failure experience.

REFERENCES

1. **H. L. Van Trees, Detection, Estimation, and Modulation Theory, John Wiley and Sons, Inc. (1968), p. 24.**
2. **G. M. Milner, et. al., " Feasibility of Vibration Monitoring of Small Rotating Machines for the Environmental Control and Life Support Systems (ECLSS) of the NASA Advanced Space Craft, " 41st Meeting of the Mechanical Failures Prevention Group, Naval Air Test Center, Patuxent River, MD, October, 1986.**

PERSPECTIVES OF MONITORING AND CONTROL OF VIBRATING STRUCTURES BY COMBINING NEW METHODS OF FAULT DETECTION WITH NEW APPROACHES OF RELIABILITY ENGINEERING

Dirk Söffker
Safety Control Engineering, University of Wuppertal
Gaußstraße 20, D-42097 Wuppertal, Germany

Uwe Kay Rakowsky
Safety of Transportation Systems, University of Wuppertal
Gaußstraße 20, D-42097 Wuppertal, Germany

Abstract: The paper introduces a perspective concept of monitoring and control of vibrating structures by combining new methods of fault detection with new approaches of reliability engineering. The complete concept is demonstrated with the mechanical example of the cracked rotor.

Key Words: Fault Detection, Fault Diagnosis, Monitoring, Reliability Engineering, Reliability Control, Safety Engineering

INTRODUCTION: Reliability and safety aspects are becoming more and more important due to higher quality requirements, complicated connected processes and structures. As examples of vibrating structures to be monitored / controled and diagnosted, the following application fields are given: turbines of power plants or modern aircrafts, controled wings of advanced materials for new large aircraft concepts, truss structures of space stations, and helicopter cabines and blades.
The monitoring and control approaches commonly used for vibrating structures are based on signal analysis methods. In case of faults, system changes etc., decisions have to be made concerning the further operations of the structures, related to questions of stopping the machine, repair and maintenance strategies, and influencing the system not only related to the goal of optimizing the vibrational behavior, but also of optimizing the reliability characteristics. The two important knowledge fields which determines this process are:
• knowledge about qualitative and quantitative effects of the faults
or in general of system changes, and the
• knowledge about the reliability characteristics of the components and the system in the fault-free and the faulty state. This includes the knowledge about the past stress of the system and the detailed knowledge about the effects of the actual changes / faults related to reliability and safety questions.
Today these decisions are typically handmade and not solved in a problem adequate way. Contemporary reliability engineering offers three ways to improve the performance of a

system relevant for safety:
- decrease the number of components,
- increase the reliability properties of selected components, and the
- use of redundancy concepts for selected components or modules.

To decide which of these concepts is the best one, the design-team provides a detailed model of the system during the early construction phase. The model is supposed to be as close to reality as possible, but it is often based on assessments about the behavior of the system in the operation phase. A lot of estimations have to be made about external influences and their influence on reliability characteristics.

The idea of this paper is to overcome these person- and problem- related decision problems.

The problem is considered from a system theoretic point of view. The idea is to get a better and objective inner view to the system and its changes. Assuming that knowledge concerning the consequences of these changes is available, the actual state of the system can be expressed by reliability characteristics. These relations can be also used vice versa for control approaches.

The paper is organized as follows: This section introduces into the problems of causal fault detection and related operating decisions. Getting a better and more objective inner view into the structure some requirements to Fault Detection and Isolation (FDI-) Approaches are formulated from the point of view of the closed loop consideration of section 2. In section 3 the interaction of the elements of the proposed Safety and Reliability Control Engineering (SRCE-) approach are declared. Section 4 demonstrates the concept from a principal point of view. Therefore the classical mechanical example of a cracked rotor will be used as a principal example. Section 5 briefly shows in which application fields the proposed ideas also can be used.

SAFETY AND RELIABILITY CONTROL ENGINEERING - CONCEPT

The new Safety and Reliability Control Engineering - Concept is given in fig. 1 here firstly in detail. The system to be considered will be understand as an input-output system. In this way descriptions (e.g. differential equations) are used, which describe the time behavior of the outputs (here: measurements - signal data) and the inputs (here: operating parameter and/or control inputs).

The task of the Reliability Evaluation Modul (REM) is the real-time calculation of the reliability characteristics of the system to be observed. Therefore only the inputs, the outputs, and advanced system informations can be used. Advanced system informations include the knowledge of the influences of changes and the ability of the modul to ensure correct informations concerning to fault changes. This implies robust approaches to ensure these properties for the use for - due to changes - variable systems. The output of this module are reliability characteristics. Problem dependent this can be

- a failure rate

$$h : t \rightarrow h(t), \tag{1}$$

- a (failure) probability density function f with

$$f(t) = h(t)e^{-\int h(t)dt}, \tag{2}$$

- or its cumulated density function F

$$F(t) = 1 - e^{-\int h(t)dt} \tag{3}$$

which is the time dependent failure probability. In most applications it would be convenient to use the failure rate.

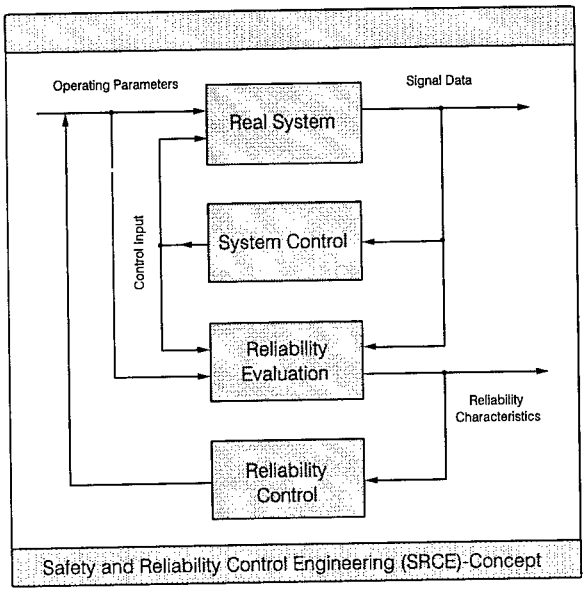


Fig. 1: Safety and Reliability Control Engineering Concept

The output of the Reliability Evaluation Module represents the actual information concerning the reliability characteristics. A simple control strategy can be easily build up by observing the output of the Reliability Evaluation Module. Related to threshold values

- alarms can be given starting intensive supervision / controling,
- or the system can be stopped immediately.

Because of the importance of this modul, this task is decribed in /4/ seperately in detail.

The idea of the SRCE-Concept is to establish a closed loop approach with the goal of controling reliability characteristics. If knowlegde is available concerning the relation

Operating Parameter $\rightarrow$ Actual System Behavior $\rightarrow$ Reliability Characteristics

these causal direction has to be reversed.

If the system structure- or parameter changes due to faults, all inputs related to the goal of controling the reliability characteristics can be used as control inputs of the SRCE-control loop

$$\text{Reliability Characteristics} \begin{array}{l} \leftarrow \text{ System Configuration / Reconfiguration} \\ \leftarrow \text{ Operating Parameters} \\ \leftarrow \text{ Maintenance and Repair Strategies.} \end{array}$$

In this way problems appear concerning the realization of the closed loop. Because of the fact that this control loop approach is not a pure technical control approach, a part of the necessary connections can not be given yet. Therefore an optimal solution will

be the use of qualitative or quantitative and unique mathematical equations, which is prefered here.
In the following some principal solutions are introduced for understanding the SRCE-idea.
i) In this way the relation $rc\,(h, f, F) = function(op, sp)$ between the set rc of reliability characteristics $rc\,(h, f, F)$ and the problem dependent operating parameters op acting on the system with the system properties sp can be used by the inverse relationship

$$op = function^{-1}(rc\,(h, f, F)_{des.}, sp)\ . \tag{4}$$

ii) If system reconfiguration strategies are also included into the SRCE-concept more difficult strategies have to be used, to solve the appearing problems.
iii) The simplest way is to change the system behavior by (inner technical) control. Due to system changes the fault-free dynamical system A_{ff} changes to the faulty-one A_f. The real-time Reliability Evaluation Module calculate the corresponding change of the rc-values as

$$\text{Actual system behavior } A_{ff} \overset{\text{due to system changes}}{\longrightarrow} A_f \tag{5}$$

$$\text{Related reliability characteristic } rc_{ff} \longrightarrow rc_f\ . \tag{6}$$

The task of an (inner technical) control loop may be to adapt the control law, building up the new system $\tilde{A}_f$ to establish minimum or desired reliability characteristic $rc_{des.}$. Assuming a state space approach, the control law KB is

$$KB = -(A_f(rc_f) - \tilde{A_{ff}}(rc_{des.})). \tag{7}$$

Because of this indirect relation related to the controled value rc the correctness of the relation
Actual System Behavior $\leftrightarrow$ Reliability Characteristics
becomes to be the most important part of the SRCE-Concept.
iv) In opposite to the above mentioned inner control loop, system reconfiguration approaches lead to much more difficult problems. From a principal point of view (also using a state space representation) the problem appears as follows: The system will be modeled using the description A_{ff}. In contrast to (iii) here it is assumed that this complete model description is not problem adequate available, or the desired rc - parameters can not / should not established by this way. In spite of this disadvantage it is furthermore assumed that submodel descriptions A_{i-ff} are available and the configuration concerning the interactions built up using the relationships cf are known. The knowlegde allows the calculation of the reliability characteristics $rc_{compl.}$ of the complete system $A_{compl.}$. These assumptions allow the formulation of the resulting SRCE-problem by

$$rc_{compl.} = A_{compl.}(cf(A_{i-ff}))\ . \tag{8}$$

The degrees of freedom concerning the control of the rc are given by
- configuration of cf,
- inner (technical) control loop of the elements A_i and
- combinations of both.

Also here the main task of the concept - the establishment of relations -
Actual System Behavior $\leftrightarrow$ Reliability Characteristics
becomes very important.

ELEMENTS OF THE SRCE CONCEPT:

Causal fault detection approaches for establishment the relation: measurements - reliability:

The proposed concept controling the reliability characteristic of a system will be applied to mechanical structures. Here the main tasks are

- the determination of suitable strain characteristics describing mechanical system states which allow to establish connections to the reliability characteristics and
- building up control laws.

Assuming that beside the measurements, system informations (e.g. input - output relations as differential equations or linguistic (fuzzy) input - output relations as qualitative description) are available, different paths of creating connections from the measurements to the interesting reliability characteristic values are possible, cf. fig. 2.

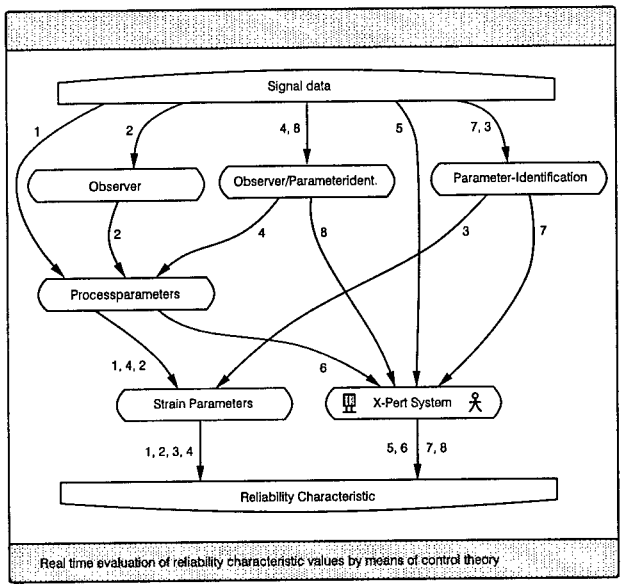


Fig. 2: Paths from Measurements to Reliability Values

Starting with the goal of the Reliability Evaluation Module - the establishment of reliability characteristics two paths are possible:

- a **physical oriented path**, and
- a **phenomenological oriented path.**

Because of the fact that the measurable signal does not content the interesting reliability expression intermediate levels are introduced, which allow the stepwise transition.

Both paths connect intermediate levels (of physical parameters (physical path) or of linguistic formulated - or in general qualitative - expressions) with the goal of building

up connections to the reliability characteristic values to be looked for. Both paths can be found using the experimental related knowledge of the life-time behavior of the structure or the estimation knowledge and experience of the reliability analysis team, cf. /4/.
The more physical oriented path uses physical intermediate levels. The most important level therefore is the level described by strain parameters. Here the reliability related term 'strain' is described by the mechanical related term 'stress', here used as mechanical stress.
In mechanical structures these kind of level is typically directly related to the reliability parameters. Therefore the classical experiments are done and expressed with series of figures of Wöhler - curves. These strain parameter level typically can not be derived directly form the measurable signals. In simple cases the different strain parameters (here: mechanical stresses) will be summarized by handmade formulas with practical proved coefficients.
Different situations appear:
• The measurable values represent the process parameters directly and only has to be combined using non-variant system parameters to get the strain parameter, typically by some algebraic equations (Path 1).
• Inner states of mechanical structures often are not measurable. Parameter changes also can not be measured. Observer techniques, parameter identification techniques or combined approaches, can be used to get an inner view into the system using inner states (e.g. to the process parameters to be looked for) and other relations modeled using system informations (Paths 2,3,4). Using measurements and observer-based estimated inner states, process parameters can be calculated. Remarks related to the use of such techniques are given in /8/. The important criteria of these techniques is the ability of building up causal connections between measurements and outputs of these advanced modules, to get a real view to the system changes.
The principal ideas using advanced control techniques for monitoring and supervision are illustrated briefly for parameter identification approaches and observers in the sequel.
Real time calculation of Reliability Characteristics: This important part of the concept is declared in detail in /4/.
Reliability control approach: Core of the control approach is the idea of controling not a physical value of the system, but the reliability characteristic of the system as an indirect value like a quality item.
This indirect parameter represents the comparison of the structure to be observed, related to the experience of a collective of identical structures (past experience), or the assessment with this structure (estimation) respectively.
The SRCE-control approach should be distinguished clearly from classical control approaches. A schematic comparison is given in fig. 3 as follows. Using the reliability control ideas introduced in section II an control loop will be established. Because of the non-existence of the necessary relationships at this stage of development of the SRCE-concept the control loop scheme can be only declared from a principal point of view.
For mechanical structures the typical strain parameter seems to be the mechanical stress σ. The stress σ results by relations of

$$\sigma = \sigma(\text{Forces, Torques Geometry}) \tag{9}$$

and is - concerning to reliability aspects - related to maximal tolerable mechanical stress of the structure

$$\sigma_{rc,max,tol} = \sigma_{rc,max,tol}(\quad \text{Forces, Moments, Load History, Material Properties,} \quad (10)$$
$$\text{Temperature, Environment Influences)}.$$

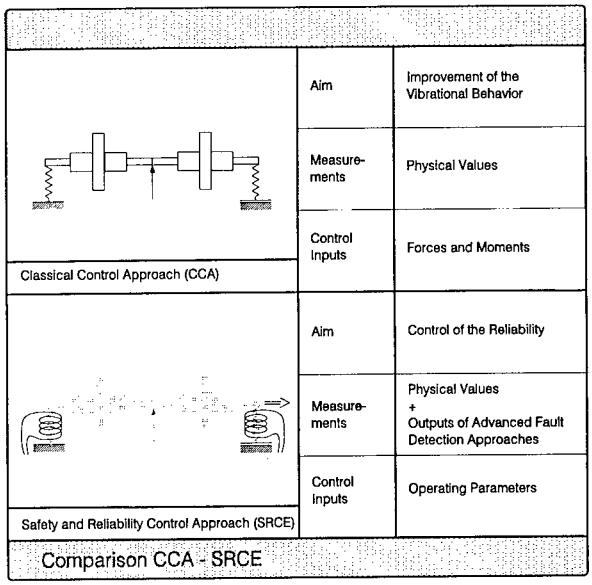


Fig. 3: Comparison CCA - SRCE approach (cracked rotor example)

The connection between the load history of mechanical stresses and strains and the failure rate is assumed to be described by the material depended Wöhler curve.

The mechanical stress σ directly depends on the static loads l_s, the dynamics loads l_d, the circular and rectangular cross sections W_p, W and the area A by

$$\sigma = \sigma(l_s, l_d, W_p, W, A) \; : \; rc\,, \tag{11}$$

which corresponds to a set of reliability characteristics rc. The reliability related stress $rc_{tol} \; : \; \sigma_{tol}$ depends additionally from the load histories $l_s(t), l_d(t)$, from the time t because of material ageing processes, from the preload history $l_s(t-\tau), l_d(t-\tau)$, the temperature ϑ as

$$\sigma_{tol} = \sigma(l_s, l_d, W_p, W, A, t, l_s(t-\tau), l_d(t-\tau), \vartheta). \; : \; rc_{tol}\,, \tag{12}$$

The idea given in eq. (5) can be used for control. Typically only a few parameters can be used for SRCE-control (e.g. the dynamical load by decreasing the power, the angular velocity of rotating systems etc.) depending on the problem and the structure itself.

The control loop is closed by the system changes itself. System changes are detected by the REM and analyzed concerning the reliability. The loss of reliability

$$\Delta rc = rc_f - rc_{tol} \tag{13}$$

has to be canceled by the SRCE-control loop.
In general the problem can be formulated as the compensation of the reliability-characteristics-decrease $\Delta rc = rc_f - rc_{tol.}$ caused by system changes due to

$$\Delta rc = rc_f(W, t, A, \vartheta) - rc_{tol.} = \Delta rc_{SRCE}(l_s, l_d, l_s(t-\tau), l_d(t-\tau), W, A, \vartheta). \quad (14)$$

This includes that system changes leading to cracks etc. resulting in changes of W, A or because of material ageing properties t, ϑ and can be determined. On the other hand Δrc_{SRCE} can use repair strategies to improve geometry properties W, A or in general operating parameters l_s, l_d, or use some known facts of improving material properties by intelligent preloading strategies using $l_s(t-\tau), l_d(t-\tau)$.

SRCE - EXAMPLE: CRACKED LAVAL ROTOR: The task of the SRCE-example of the cracked rotor is to control the reliability characteristic of the failure rate of the laval rotor similar example given in fig. 4.

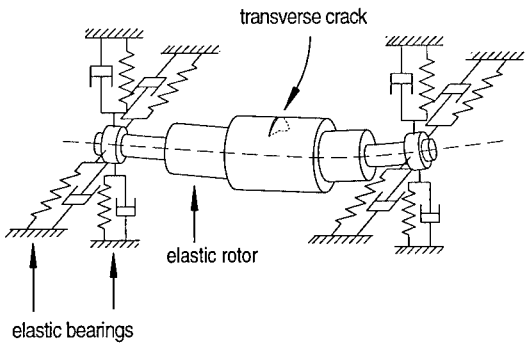


Fig. 4: Principal example of the cracked rotor
Firstly the classical supervision problem will be declared briefly. In the following the new SRCE-concept ist principally applied to this practical problem.
Classical supervision concept: During the stationary operation, measurements are taken from the bearings. Using signal analysis methods the frequency behavior will be monitored very well. Changes of the mechanical structure are leading to signatures of the spectrums, to changes concerning the amplitude of the observed vibrations etc. . Using human or artificial pattern recognition, comparing the vibration amplitudes with maximum allowed values, changes can be detected. Because of the fact, that the causal relation between these parameters and the physical reason is not unique, a very difficult decision - problem for the operating staff appears, which depends on the knowledge and experience of and with the system and also of economical and psychlogical constellations. In the - from a statistical point of view - rare case of a transverse crack, different phenomena - viewing the vibrational behavior - appears. Therefore different diagnosis philosophies exist. (e.g. /3,9/). Independent from the unsafe and ambiguous decision, no statement can be given about the depth of the crack and the real reducing of the technical functionality and the reliability. To be sure, the operation has to be stopped, the machine should be taken apart. Related to the viewed fault the operation is continued with lower strain parameters up to the next repair or inspection date, or the system

has to be repaired immediately. The complete process is strongly handmade and can be optimized in several cases using available techniques.
As a common goal of the application of advanced techniques the introduced SRCE-concept can be used.

SRCE - Concept: The relevant RME-modul for crack detection consist of two parts. The first part is the Proportional-Integral Observer (PIO) /7/, which is applied to the cracked rotor /6/. Here (with the same measurement information as used for the classical supervision concepts - the displacements of the bearings -) and additionally the mechanical description of the fault-free system, the effects of the crack are estimated directly at the crack location. This means by using the system information A_{ff} and simple measurements y, the PIO gives the crack depth as output, cf. fig. 5.

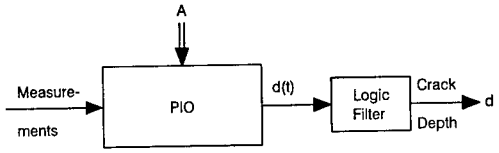


Fig. 5: Scheme of the PIO as RME-element

The crack depth d directly reduces the mechanical properties of the cross sections W, W_p directly. The explizit knowledge are given in the works of Mayes and Davies /2/. Here these analytical - geometry and material properties considering - algebraic formulas are used to give the relation crack depth - loss of cross sections as shown in fig. 6.

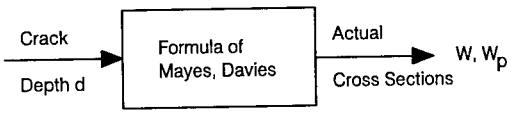


Fig. 6: Establishing the relation crack depth - cross sections

Assuming that the mechanical stresses of the simple structure of the Laval rotor are described by

$$\sigma_b = \frac{M_b}{W} = \frac{mg\varrho^2 L}{2W} \text{ and} \tag{15}$$

$$\tau = \frac{M_t}{W_p} = \frac{P}{2\pi n W_p} \tag{16}$$

with the bending moment M_b, the torsional moment M_t, the mass m, the gravity g, the density ϱ, the length L of the rotor, the power P, the revolutions n and the cross sections W, W_p shown in fig. 7.

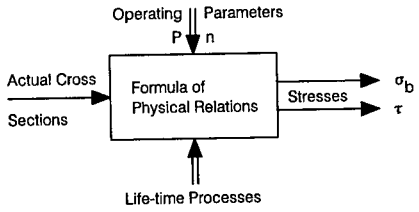


Fig. 7: Physical relations leading to strain parameters (here: mechanical stresses)

In contrast to a new approach illustrating advanced reliability modeling strategies /4/, here the old practically proved formula combining mechanical stress parameters is used, resulting to one compatible mechanical stress value

$$\sigma_c = \sqrt{\sigma_b^2 + 3(\alpha_o \tau)^2} \tag{17}$$

with the practically found parameter $\alpha_o = 0.63$ for turbine steel, cf. fig. 8.

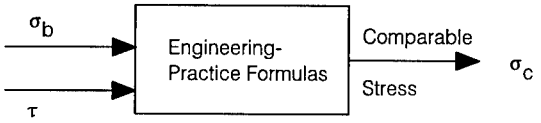


Fig. 8: How to combine two strain parameters to one

This parameter can be compared with the values found by numerous experiments combining the mechanical stress values, the number of stress changes and the failure rate: called Wöhler - diagrams, cf. fig. 9.

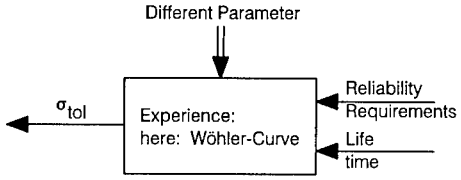


Fig. 9: Simple Wöhler diagram

The formula describing the time dependent strength properties with a failure rate of 1% is

$$\sigma_{rc}(N) = \left(\frac{\lg \frac{\sigma_F}{\sigma_D}}{\lg \frac{N_F}{N_D}} N + \lg \sigma_F - \lg N_F \left(\frac{\lg \frac{\sigma_F}{\sigma_D}}{\lg \frac{N_F}{N_D}}\right)\right) ind[N_D - N] + \lg \sigma_D \; ind[N - N_D] \tag{18}$$

$$ind[arg] = \{\ 1 \text{ for } \arg \geq 0;\ 0 \text{ for } \arg < 0\ \} \text{ with} \tag{19}$$

$\sigma_F, N_F, \sigma_D, N_D$ as pairs of the Wöhler curve and N as the number of stress changes.

The goal of this example is to control the failure rate of 1%. Working at a lower stress level for σ_D, this means that the structure theoretically has in infinity life-time. Due to a crack at N_1 load cycles the mechanical stress increases. At load cycle N_2 the SRCE-control detect an inacceptable loss of the failure rate, because of the control goal dependent logic assumed for this example given in fig. 10.

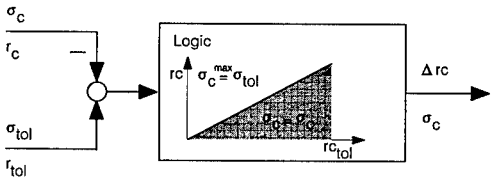


Fig. 10: Example dependent steering logic of the control unit.

Here different logic or analytic functions as control laws are possible. Potential field oriented approaches with the aim, that the max. rc-bound can not reached, are also of interest.

If the control unit works, in the case of setting restrictions to the allowed comparible stress $\sigma_{tol} = \sigma_{c,des.}$ the SRCE-control laws can be calculated using the inversion of eqn. (15-16) by

$$P_{des.} = \sqrt{\frac{2\Pi n W_p}{3\alpha_o^2}(\sigma_{c,des.}^2 - \frac{m^2 g^2 \varrho^2 L^2}{2W})}, \tag{20}$$

given in fig. 11.

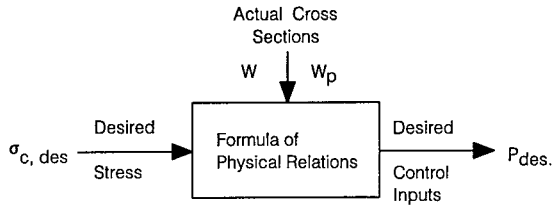


Fig. 11: Calculation of the desired control input

Using the operating parameter of the power P the SRCE-loop for this simple example is closed. The complete loop is given in fig. 12.

PERSPECTIVES OF THE SRCE CONCEPT: The idea of the SRCE-concept, the optimization of the life-time operation of systems by integration of

- advanced fault detection and isolation schemes for monitoring,
- the reliability evaluation module for calculation of reliability relevant values, and
- the control unit integrating inner loop control, maintenance- and repair strategies

into a mainly automated concept, can be also applied to other problems, in which the repair and maintenance strategy are more relevant, e.g. automated manufactoring processes, or in which reconfiguration approaches are important, e.g. safety relevant industrial processes. An advanced example is the space station concept with automated repair-control.

CONCLUDING REMARKS: This paper introduces a perspective concept of monitoring and control of vibrating structures by combining new methods of fault detection with new approaches of reliability engineering. The concept of the introduced Safety and Reliability Control Engineering (SRCE-) approach is based on advanced modules of control and reliability engineering. Fault detection and isolation approaches, which allow causal fault detection, describe the actual system state and allow a view into the system. These informations are used by the Reliability Evaluation Module (REM) to calculate reliability characteristics.

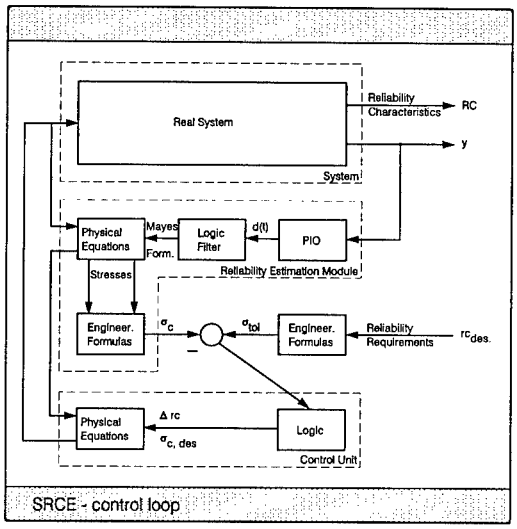


Fig. 12: Complete SRCE-loop applied to the cracked rotor example

The control loop can be established by turning back the knowledge needed for the real-time reliability determination. The complete concept is applied to the mechanical example of the cracked rotor.

REFERENCES:

/1/ Isermann, R., 1988, "Identifikation dynamischer Systeme I,II", Springer.

/2/ Mayes, I.W.; Davies, W.G, 1980, " A method of calculating the vibrational behavior of coupled rotating shafts containing a transverse crack", *Journal of Vibration, Acoustics, Stress and Reliability in Design*, Vol. 106, pp. 139-145.

/3/ Muszynska, A., 1982, "Shaft crack detection", *Proc. 7th Maschinery Dynamics Seminar*, Edmonton.

/4/ Rakowsky, U.K.; Söffker, D., 1997a, "Real time reliability evaluation of vibrating mechanical structures", submitted to: *12th ASME Conference of Reliability, Stress Analysis and Failure Prevention.* Virginia Beach, 1997, accepted.

/5/ Rakowsky, U.K.; Söffker, D., 1997b, "Safety and Reliabibility Control Engineering - Part I: Real Time Reliability Evaluation", in preparation.

/6/ Söffker, D.; Bajkowsky, J.; Müller, P.C., 1993, "Detection of cracks in turborotors - a new observer based method", *Trans. ASME - J. Dyn. Sys. Meas. Control*, Vol. 115, pp. 518-524.

/7/ Söffker, D., 1995, "A new model based tool for fault detection and isolation in machine- and rotordynamics", *Proc. 11th ASME Conf. Reliability, Stress Analysis, and Failure Prevention*, Boston, USA.

/8/ Söffker, D., 1997, "Robust fault detection of large vibrating structures by means of control theory - some principal remarks", submitted to: *12th ASME Conference of Reliability, Stress Analysis and Failure Prevention.* Virginia Beach, 1997, accepted.

/9/ Wauer, J., 1990, "On the dynamics of cracked rotor: a literature survey", *Applied Mechanics Reviews*, 43, pp. 13-17.

MOTOR SYSTEMS MANAGEMENT

John D. Kueck

Oak Ridge National Laboratory
P.O. Box 2009
Oak Ridge, TN 37830

Abstract: Motor efficiency and lifetime are often compromised when motors are applied without an understanding of motor rating and the requirements of the driven equipment. Simple measurements may be used to estimate a motor's efficiency and to perform cost benefit analysis for motor replacement. Motor Master Plus is an invaluable tool for this estimation and analysis. Variable speed drives can provide truly remarkable energy savings, and their application and some potential concerns are discussed.

A study of a data base of motor failures has provided some interesting trends. Motor failures tend to be predominantly bearing related, with stator failure second, and other failures relatively insignificant. Incipient bearing failures are generally best detected with vibration analysis, but there are some very interesting new technologies for detecting stator condition on line. Some of the conventional stator checks, such as a megger test, are sometimes of little value.

An electrical distribution system tune up can pay big dividends in ensuring optimum motor performance and reliability. Measurements can be made with relatively inexpensive instruments, and major problems are often lurking undetected.

Key Words: Condition Monitoring; Diagnostics; Efficiency; Failure; Motors

INTRODUCTION: There is presently a revolution occurring in the performance analysis and diagnostics of motors. There are a number of diagnostic tools now on the market which can evaluate motor efficiency while the motor is operating in its actual service condition, and tools which can diagnose motor problems such as cracks in the rotor bars or degraded bearings while the motor is operating. In fact, the best time to make these diagnoses is when the motor is operating and under heat and mechanical stress, and not several hours after it has been de-energized and is at room temperature. These advances in diagnostics should result in improved motor lifetime and efficiency, as well as improved operating efficiency. The revolution in diagnostics now allows an engineer to easily monitor the flow of power through the motor driven system, and to determine where unacceptable wastes may be occurring, whether they are in an improperly applied motor, a throttled control valve or damper, or due to an electrical distribution system problem, such as phase unbalance.

Motor Diagnostics and Reliability: Motor diagnostics are now to a large extent computer based data analysis devices that analyze a large number of data points in sophisticated mathematical programs to detect problems in an low intrusion manner, with a minimum of downtime for the motor. Knowledge of motor condition and performance can lead to vastly improved reliability, as motors can be scheduled for preventive maintenance prior to failure, and service conditions which may be leading to motor degradation can be corrected.

Motor Reliability Considerations: NEMA design induction motors are designed to operate in clearly defined service conditions, as specified in NEMA MG1 (Ref. 1). Operation of the motor with a 10 percent overvoltage or undervoltage at rated load may cause an increase in heating of the motor, which will, over time, result in a degradation of the insulation of the motor. Another common problem is operation of the motor with a voltage unbalance. Operation of the motor with a voltage unbalance produces a magnetic field in the motor which rotates in the opposite direction to the rotation of the rotor, producing heat, vibration and greater slip in the motor. NEMA provides a table of suggested motor derating for voltage unbalance, and for an unbalance of two percent, NEMA recommends a derating factor of five percent. Voltage unbalance is a common problem in older plants where loads have been added to the distribution system over the years without checking to ensure that voltage were kept balanced, and the author has been requested to consult on premature motor failures that were caused by a voltage unbalance when the motor was operated at full load continuously. It should be noted that no plant is going to have perfect voltage conditions, and some 1.0 Service Factor motors are now being manufactured with a manufacturer's recommendation that they not be operated at full load continuously, or premature failure may result.

Harmonic distortion in the voltage waveform can also cause motor heating which will result in insulation degradation, but for motors powered directly from the bus, it is unlikely that harmonic distortion problems will be noticed first in the motor. NEMA (Ref. 1) recommends a motor derating factor of 5 percent when the harmonic voltage factor reaches .06, using their definition. This is a severe level of distortion, and it is the author's suspicion that problems with computer failure, overheated neutrals etc., would have occurred long before premature motor failures are noted when the problem is purely voltage distortion on the bus voltage.

Other environmental factors which have an effect on motor lifetime include ambient temperatures above 40 or below 0 degrees C, altitude above 3300 feet, or obstructions to the flow of motor cooling air. Reference 1 provides a list of abnormal service conditions which can reduce the life of the motor, these include: combustible or abrasive dusts, lint, etc. in the air, salt air, abnormal axial loading, frequent starting, and many others.

Motor Failure Categories: The IEEE recently completed a study of failures in large motors (Ref. 2). The study found that for induction motors above 200 hp, roughly 50% of all failures were in the bearings, and 25% of all failures were in the windings, and the rest were scattered among the rotor, shaft, external devices, or not specified. The major contributors for the bearing failures were listed as high vibration, poor lubrication, and normal deterioration with age. The major initiators for the winding failures were: overheating, insulation breakdown, mechanical breakage and electrical fault. For induction motors above 200 hp, rotor failures accounted for only 2.5% of the total population. This is regrettable in a sense, because rotor failures are the easiest to detect using non intrusive Motor Current Signature Analysis (See Below). There are a number of laboratories, including the author's, working on diagnostics that can be used on line for assessing winding condition. In the author's opinion, bearing condition is best evaluated using vibration waveform analysis.

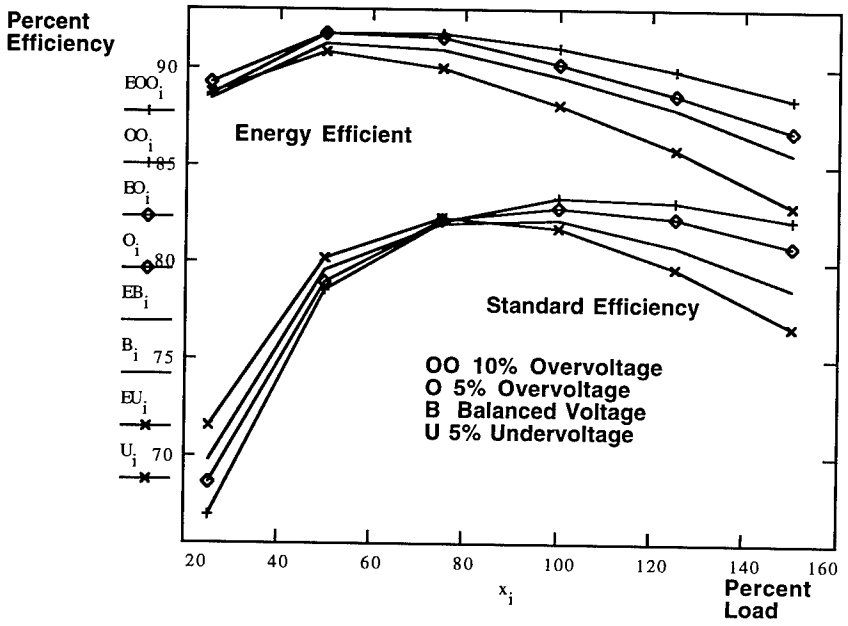
Conventional Motor Diagnostics: Conventional diagnostics include such tests as: megger and polarization index test, ac and dc high potential test, growler tests, surge tests, inductive imbalance tests, winding bridge test. These tests are all performed with the motor de-energized, and this is a significant disadvantage. When the motor is operating, it is under thermal and mechanical stresses that can cause insulation problems. When the motor has cooled to room temperature, these problems can disappear. The author has had

personal experience with a motor that was overheated to the point where its rotor had melted and one end ring was in pieces in the bottom of the end bell, but after it was allowed to cool, the motor was still passing its winding megger and resistance bridge test. Many maintenance engineers are reluctant to use high potential tests because they are concerned that they ma damage the motor insulation. Certainly the specialized tests, such as growler tests, are invaluable for detecting specific problems such as cracked rotor bars, and surge testing is very effective for evaluating stator condition on large equipment, but in general, these tests are intrusive; they require shutdown and in some cases disassembly of the motor, and they aren't used very much on installed equipment. They are impossible to use for predictive maintenance when a critical motor is operating in a running production line.

Motor Current Signature Analysis (MCSA): This diagnostic technique was invented at the Oak Ridge National Laboratory (ORNL) in 1987. This technique uses a clamp on Current Transformer to acquire the motor current waveform. This waveform is demodulated and filtered before an FFT is performed. The result is a low frequency plot of the energy in the various frequencies. Using this technology, the slip frequency of the motor (the frequency of the current in the rotor) can be clearly identified. The amplitude of the slip frequency is related to the number of broken rotor bars, and this tool is an excellent rotor condition diagnostic. MCSA is also quite valuable in detecting problems in the driven equipment because an induction motor is an excellent low frequency transducer, and many problems in gear trains, fan belts, etc. are clearly manifested by a low frequency signature in the motor current waveform. ORNL has licensed this technology to a number of diagnostic equipment vendors. The technology has now matured and is being used also in the high frequency domain and is quite useful for detection of unique faults in a variety of specialized motor and driven equipment.

Motor Diagnostics Research: ORNL, as a partner in the Oak Ridge Centers for Manufacturing Technology, now has an Energy Conservation Center with significant capability in the analysis of 480 volt, three phase induction motors up to 100 horsepower. The Center has an eddy current water cooled dynamometer, and extensive computer based data acquisition and analysis capability. The Center will have the capability to test to 500 hp soon. The Center also has a 3,000 gpm flow loop which is highly modular and can be assembled in many different configurations for testing pumps, flow transmitters, flow orifices, etc. Again, the flow loop also has extensive computer based data acquisition and analysis capabilities. A typical data acquisition scheme is to sample as many as eight channels at a frequency of 5,000 Hz for 10 seconds. The data acquisition scheme is LabVIEW ™ based, and many initial analyses, such as motor efficiency and CT phase shift correction, are programmed into the Virtual Instruments acquiring the data.

Motor Efficiency: In general, it is cost effective to replace standard efficiency motors with energy efficient motors. Energy efficient motors typically provide higher efficiency over the entire operating range. The figure below shows the results of a test of two 50 hp NEMA Design B motors from the same manufacturer, one standard efficiency, and one a premium energy efficient motor. It may be seen that regardless of the voltage condition or load, the energy efficient motor is always more efficient.



The Department of Energy has developed a program called Motor Challenge whose purpose is to illustrate the cost benefits of optimizing the efficiency of motor driven systems. There is no charge to join Motor Challenge, and the program supplies a wealth of publications and training for optimizing motor driven system efficiency, as well as some very useful software. One of the software programs available through Motor Challenge is called Motor Master Plus (Ref. 3). This menu driven program provides data base of more than 10,000 NEMA Design B motors, including performance and price information. The program contains a field data module which holds all the measurements one would normally take in a motor field survey. The program has the unique capability of calculating a motor's efficiency and load factor based on a measurement of speed or current or voltage, and using performance data from a group of similar motors in the data base as a basis for this efficiency and load factor calculation. Armed with the efficiency and load factor, the program can then use inputs of the local electricity cost and motor operating hours to perform an accurate cost benefit analyses of replacement motors. It can even do these in batch modes for groups of motors in a plant. This program allows the plant maintenance engineer to easily assemble a sophisticated motor survey for his plant and to do a state of the art cost benefit analysis for replacement motors.

Reference 4 provides a survey of approximately 30 methods for estimating motor efficiency in the field, but for the simple combination of ease of use and accuracy, the Motor Master Plus method scores as one of the best. For the user who is interested in evaluating his motor in more detail, ORNL has prepared a software program called ORMEL (Oak Ridge Motor Efficiency and Load, Ref. 4). This screen driven interactive program develops an equivalent circuit based on the motor nameplate data, and the user may "fine tune" the

equivalent circuit to match the specific motor being examined. The program also calculates and reports the losses of the motor - stray load loss, friction and windage, core loss, etc. For the user who wishes to assess the operation of his motor under a range of voltage conditions, ORNL has developed a program called IMPET (Induction Motor Performance Evaluation Tool) that allows the user to match the motor to a number of measured parameters and ensure that the equivalent circuit is a valid predictor of motor performance. (Ref. 4) This program is also interactive, screen driven and user friendly, and is available for licensing.

Adjustable Speed Drives (ASDs):

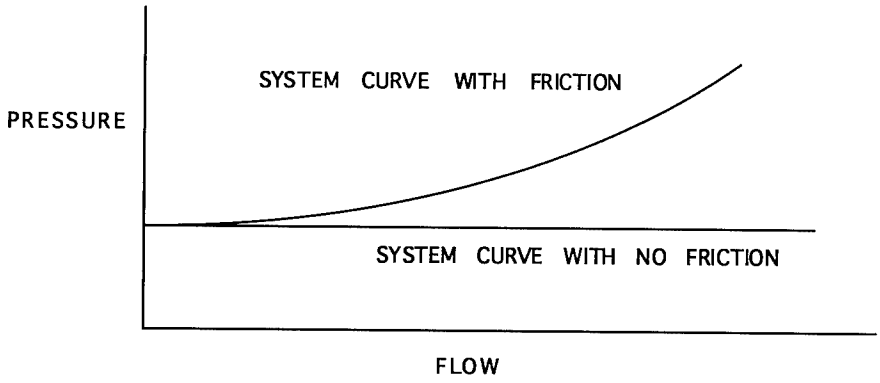
When a motor is operated at a low load condition, the user may also want to have the motor slow down, as well as providing less torque. Torque is roughly proportional to the square of the voltage. Just reducing a motor's voltage to reduce its loading could easily result in a condition where it does not have enough torque, and it would stall. An adjustable speed drive varies both the frequency and voltage supplied to the motor, thus allowing the motor to slow down significantly at low loads and to operate efficiently at a wide range of loads. The simple ASD concept is that the 60 Hz AC is converted to DC in the rectifier section, and then the DC is converted to adjustable frequency AC in the inverter section, and power is supplied to the motor at the voltage and frequency the motor needs for its present load.

There are three basic classifications of ASDs - Variable Voltage Inverter (VVI), Current Source Inverter (CSI) and Pulse Width Modulated (PWM). In the VVI, the rectifier converts the supplied AC to variable voltage DC. A large capacitor is connected in parallel with the rectifier output. The voltage is set to the needed motor voltage. The inverter converts the DC to the required frequency in six or twelve steps of current per cycle. In the CSI, the rectifier converts the supplied AC to a fixed voltage DC. A large inductor is connected in series with the rectifier output. The inverter converts the DC to a variable frequency and variable voltage AC. CSI drives have a high hp range, typically up to 5,000 hp. In a PWM, the rectifier provides a constant DC voltage. The inverter supplies high frequency voltage pulses to the motor; the width of the pulse is adjusted to provide the needed voltage and frequency. Most ASDs sold today are PWMs.

Because of the high frequency of the voltage pulses to the motor, the PWM can subject the motor to voltage stresses. In the case of very long feeder cables between the motor and drive, standing voltage waves can develop which can damage winding insulation. The ASD supplier should always be consulted regarding the motor selection and cable length to ensure there won't be problems with the application. Some ASD suppliers provide free application software. The Electric Power Research Institute has developed application software for Adjustable Speed Drives. (Ref. 5)

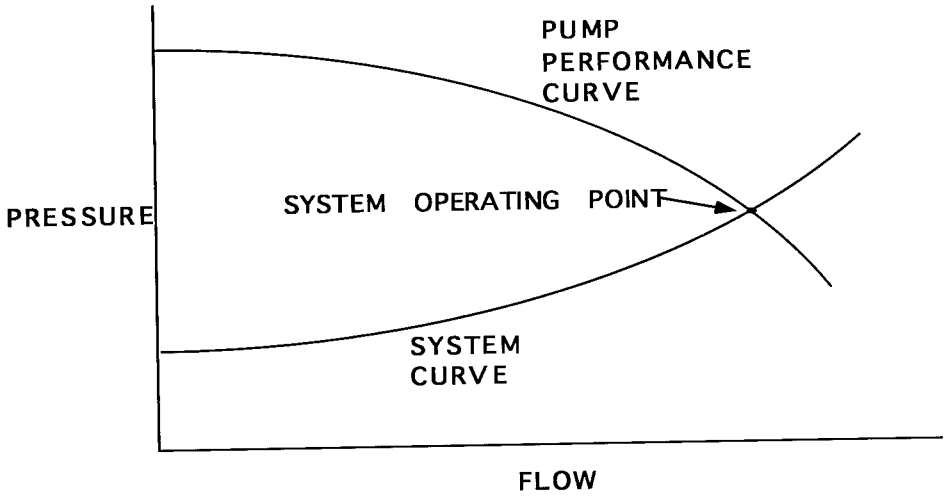
ASD Applications and Fluid Flow System Basics: The first step in determining whether an ASD would be useful for a specific application is to make a few measurements and characterize the duty cycle. Find the pressures and flows the system operates at, and the durations of each. Find the motor loads for each condition. A good ASD application has a high percent throttling and high annual operating hours. Ensure that the existing equipment is not just oversized. An ASD can offer a range of very low to greater than rated speed. If only a few operating points are required, a multispeed motor may be a better choice. Another advantage of an ASD is the opportunity for system tuning. Normally a pump and motor are each specified with at least a 5% margin, and there is at least a 10% margin in the system design, for a total of 20%. This power is usually wasted across a throttling valve or damper. With an ASD, the valve or damper can be opened fully, and the system can be tuned for the best efficiency.

In a fluid flow system, the system curve is the plot of the pressure required to push a range of fluid flow. For a frictionless system, this curve would be horizontal, but in real life, there is friction, and higher pressure will be needed to push increased flow.

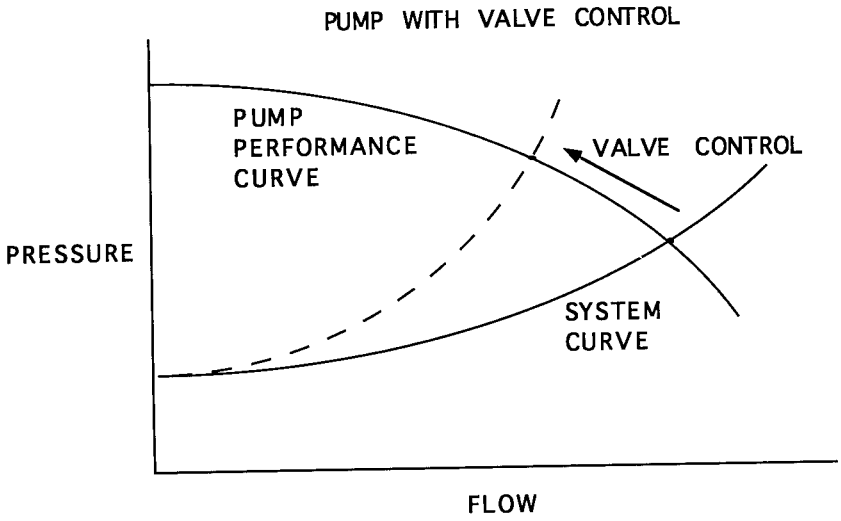


System Curve for Frictionless System and Typical System

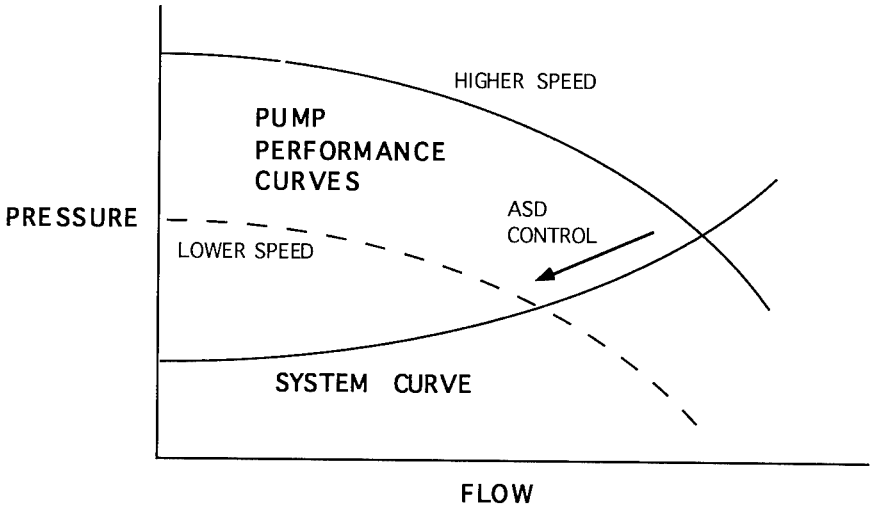
A pump or fan has a characteristic performance curve for a given speed. Usually a pump will produce more flow at a lower head. The intersection of the performance curve and the system curve is the Operating Point of the system. This is the point at which the flow and pressure will operate unless the friction changes, by opening a valve, or the performance curve changes, by changing the speed of the pump.



System Operating Point

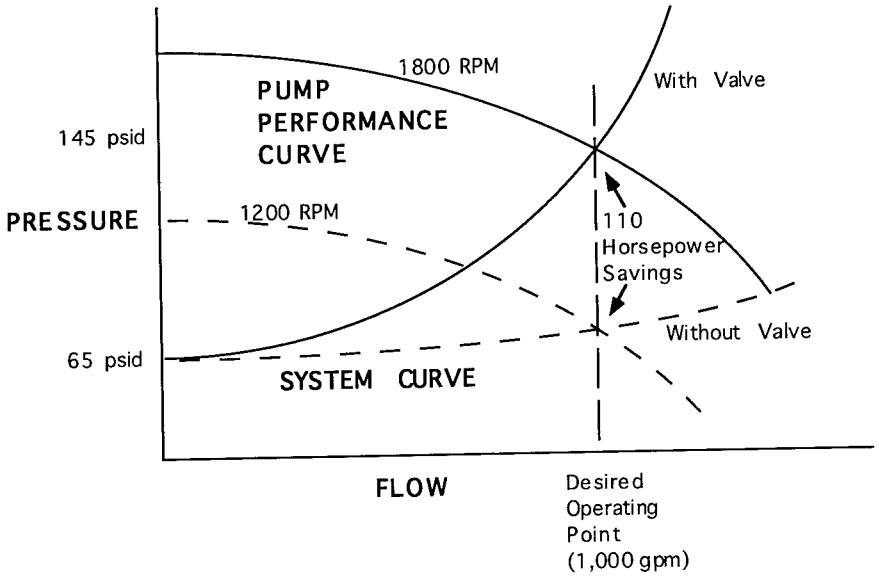


Moving the Operating Point by Changing Friction (Valve Position)



Moving the Operating Point by Changing Pump Speed

As an example of moving the operating point to improve the efficiency of a system, the following figure shows how the performance curve is shifted down by replacing an 1800 rpm motor with a 1200 rpm motor, and the system curve is also shifted down by opening the throttle valve. The difference between the old and new operating points is 110 horsepower. About one half of this savings is due to more efficient operation of the pump, and the other half is the elimination of losses at the throttle valve. In addition to the energy savings, other improvements are a significant reduction in noise form the system, and the pump seals last much longer. These results were realized by simply making a walk down of the system, taking a few measurements, and then performing some relatively simple calculations.



Old Motor: 350 Hp, 1800 rpm, running at 170 Hp
New Motor: 125 Hp, 1200 rpm, running at 60 Hp

Electrical Distribution System Tune Up: In addition to making measurements of the mechanical system, it is extremely useful to make measurements in the electrical distribution system. As mentioned above, voltage unbalance is a common problem and can drastically shorten motor life. Voltage unbalance is easy to detect, and can be caused by unevenly distributed single phase loads, high contact resistance, and possibly an open circuit somewhere in the distribution system. Under or over voltages are also common even on well designed systems and can also cause shortened motor life and reduction in efficiency. Harmonic distortion may be caused by non linear loads in a neighboring plant, and are becoming more and more of a problem to control system reliability. They are now easy to measure with one inexpensive instrument that can also measure power and voltage unbalance. A single meter can now be purchased for around $1,000.00 that measures and records the current and voltage waveforms, calculates power, and allows the user to view the current and voltage waveforms on the spot, along with harmonic analysis. It also allows the user to download the recorded waveforms onto a personal computer for further analysis. It is really worthwhile to make these measurements of voltage, current, power, pressure and flow, and evaluate the power flow through the motor driven system, because there are often surprises lurking, and typically, improving efficiency also means an improvement in reliability.

Acknowledgments: The assistance of D. A. Casada and S. L. Bunch is appreciated in developing the fluid flow basics portion of this paper and in providing the example of energy savings using a reduced speed motor.

References:

1. NEMA Standards Publication No. MG1 - 1993, Motors and Generators, Revision No. 1, December 7, 1993

2. IEEE Committee Report, P. O'Donnell, "Report of Large Motor Reliability Survey of Industrial and Commercial Installations" IEEE Transactions of Industry Applications, July August 1985.

3. Motor Master Plus, Developed for the U.S. Department of Energy by the Washington State Energy Office with Support from the Bonneville Power Administration.

4. Contact Dr. Pedro Otaduy, Oak Ridge National Laboratory 423 576 5125.

5. ASD Master, Electric Power Research Institute, Adjustable Speed Drive Evaluation Methodology and Applications Software, AP-106949.

THE VALUE OF INTEGRATING PROCESS VARIABLE MONITORING WITH VIBRATION MONITORING - TOTAL CONDITION ASSESSMENT

Gerald K. McGowan

IDAX, Inc.
5301 Robin Hood Road, Suite 134
Norfolk, Virginia 23513

Abstract: With the phenomenal increase in the capabilities of reasonably priced microprocessors and recent advances in Machinery Monitoring knowledge, there is a trend toward comprehensive on-line machinery condition assessment. The powerful combination of on-line vibration integrated with process data allows the user an incredible amount of flexibility, not only for maintenance program but for total process monitoring and control. This integration provides a tool that can greatly aid in both Maintenance and Operations in the following areas.

- Decreased cost of production
- Decreased number of costly, unscheduled shutdowns
- Increased overall plant efficiency
- Increased inter-departmental cooperation

Our purpose here is to briefly examine advances in several of the enabling technologies. We will then discuss their application to achieve an integrated system and demonstrate the benefits of such a system.

Key Words: Condition Based Maintenance; New Technologies; Performance Optimization

Introduction: We are in the business of maintenance and as a group we focus our interests and energies squarely on maintenance; Predictive, Proactive, Profit Centered, and Reliability Centered.

Each of the above maintenance philosophies is constantly and rapidly evolving. There are reports of advances in the philosophies and in the enabling technologies every month in professional publications. It is truly an exciting time and a constant challenge to keep abreast of the latest changes.

This discussion will only look at several examples of change in the following broad areas: Hardware and software, Internet/Intranet/Mobile Computing, Open Systems and Expert

Systems. After the above discussion, several examples of implementation, both current and planned, will be presented.

Hardware and Software: One of the most significant advances in the recent past is the unbelievably fast change in the capabilities of the PC. Just two years ago shipments of the 90Mhz Pentium processor backed up by 64 Mb of RAM was eagerly awaited. The extraordinary increases in performance obtained from the Pentium 60, 66 and 75 MHz processors led to predictions that the 90 MHz base PC would relegate the trusted 486 to the closet.

Today, Dual Pentium Pro processors with speeds of 200 MHz, 256 Mb RAM and a 4Gb hard drives are readily available. What is truly enabling about this technology is the price. The high end system costs the same or less than the Pentium 90 based PC of 2 years ago and the cost of additional memory has bottomed out at about $10 per megabyte.

A positive side effect of the above advancements is that those older, slower processors (the x86 family) are available in large numbers and at nearly give away prices. They provide the perfect platform for implementing local intelligent collection and processing of data. Local storage of the data is on inexpensive, hot swappable hard drives or re-writable optical storage disks. The preprocessed data is also broadcast to a data server or to an on line monitoring system over a fiber Optic LAN running at 100 MB/s.

Sensors are more intelligent, hardy and less expensive and the selection ranges from specific task to multifunction. Today, even the Energizer Bunny has disposable sensors on his product.

Utilizing intelligent local collection and processing of sensor data reduces the workload of the monitoring system. It also frees up resources of the data server, and the LAN. These resources are then available for processing more data and producing information that is packaged for specific end users.

Implementation of such a system is also less expensive than traditional point to point systems. The acquisition cost of the processor, associated hardware, and software are low but the real cost savings are realized in reduced installation time and impact on plant operation. Hookups to data points are all localized and the only point to point installation is the LAN. Cost per data point is down and will continue to fall.

Breaking the Tether: One of the most exciting areas of phenomenal growth is wireless communications and mobile computing. Spurred by demands of the business community, technologies are now in place which allow the user access to data and applications without the need for a physical tether to the network. Easy, cheap and almost universal access to the Internet allows the user to connect to virtually anything from anywhere and the maintenance community stands to benefit.

The impact of the Internet has caused major changes in the way the systems are perceived. There is now a positive trend to use the Internet and Intranets as the foundations for standardizing enterprise operations and practices. A major advantage of this move is that the source of data, whether on a local drive, on the Main Frame, or on the Internet, is irrelevant to the operator. Another positive result is that the Operating System (OS) of the data source is immaterial as it is invisible across the network.

Open Systems: Microsoft's family of Windows operating systems (Windows 3.x, Windows 95, Windows NT) and the well established UNIX family have emerged as the defacto standards for operating systems. The providers of these operating systems are well established and can be expected to push the technologies forward and provide continued support. A new attitude with respect to software is now evident.

Industry now demands standard software packages written for the dominant Operating Systems instead of many one of a kind applications on non standard, proprietary Operating Systems. They are also demanding that new systems integrate with their established and future systems. In short, they demand connectivity and open systems. As a result of all of the above, suppliers of new technologies are responding with several initiatives.

- Open Data Base Connectivity (ODBC) ... Compliant software provides assurance that data can be shared between different systems.
- Object Linking and Embedding (OLE) for Process Control. (OPC) ... An OLE based interface standard. OPC's objective to provide greater interoperability between control applications, systems and devices, and business and office applications. [1]
- Machinery Information Management Open Systems Alliance (MIMOSA) ... A group of users, instrumentation, technology and service providers. MIMOSA will deliver technology tools to make machine information system integration nearly as simple as plugging in a VCR.[2] An open exchange of information through a 'common ground' allows the customer to select best for application from the islands of compliant technology.
- Plug and Play ... (Windows 95, SP-90 Fieldbus standard) ...Automatic self integration of hardware and software.[2]
- VISA (Virtual Instrument Software Architecture) ... The I/O interface standard that will govern how all VXI plug and Play software components are built. [2]
- ISO 9000 and ISO 14000 Series... The standards, not the implementation. These offer standardization and a way to distinguish between providers in the market place.

Expert Systems: Until recently, Maintenance Expert systems relied almost exclusively on one - zero logic. Either a condition was true or it was not. Fuzzy Logic is now being incorporated into several expert systems that will soften these absolute limitations. This will allow Expert Systems to be written in 'Human' terms such as 'Nearly aligned' or 'slightly open'. Decisions based upon fuzzy logic become much more natural.

In addition to fuzzy logic, the expert systems have the potential of becoming genius systems with the application of neural networks. Neural networks will add pattern recognition and matching coupled with data interpretation and inference to the fuzzy expert for a system which will be capable of 'learning' from its history.

So What do we make of all of this? The previously cited advances in the enabling technologies and philosophies are all available now but what do we make of them and what does it all mean? Quite simply, this means that the purchaser is afforded an almost unlimited degree of freedom in selection of components which when carefully integrated will give:

- Higher performance at lower cost.
- The ability to 'tune' technology to the problem at hand by selecting 'best for application' from the technologies available.
- The ability to mix and match 'best for' providers with integration assured.
- The ability to grow with technology.
- The ability to give legacy systems a technology face-lift.
- The assurance of continuing support.
- Expansion into areas not previously possible.

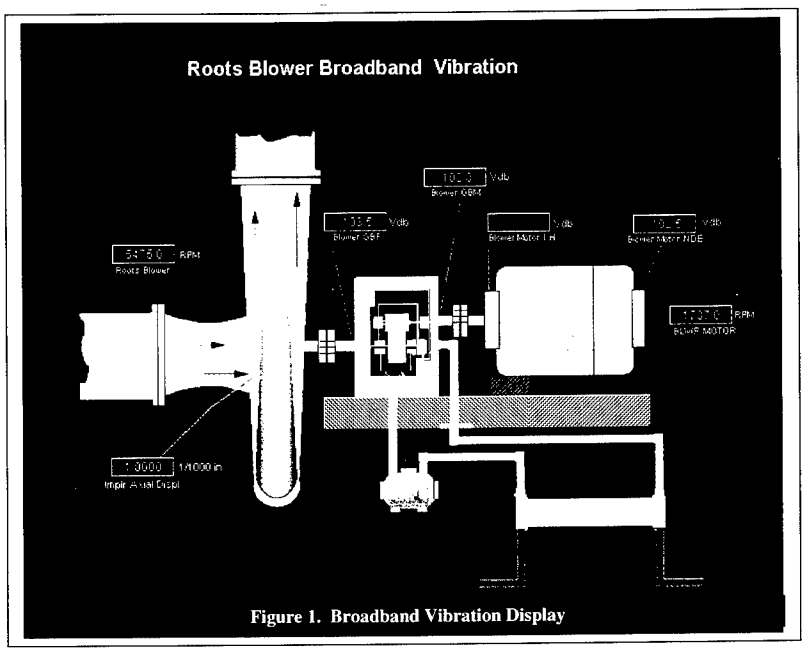


Figure 1. Broadband Vibration Display

Demonstration Case: Figure 1 is a graphical representation of a Roots Blower and its subsystems. The Blower is used as the source of heated steam and is critical to the operation of the entire plant. A major cause of damage in the past has been surge in which system conditions allow the reverse flow of steam through the blower. The customer is anxious to move to an on line monitoring system and would like to explore possibilities using the blower as a test case. No sensors are currently installed.

Maintenance Point of View with On-line Vibration: Vibration analysis is, and will always be a valuable and important part of the maintenance toolbox. It is typically non-invasive, well understood and widely accepted. It traditionally serves as a stand alone intenance technology or as the cornerstone of an integrated maintenance program.

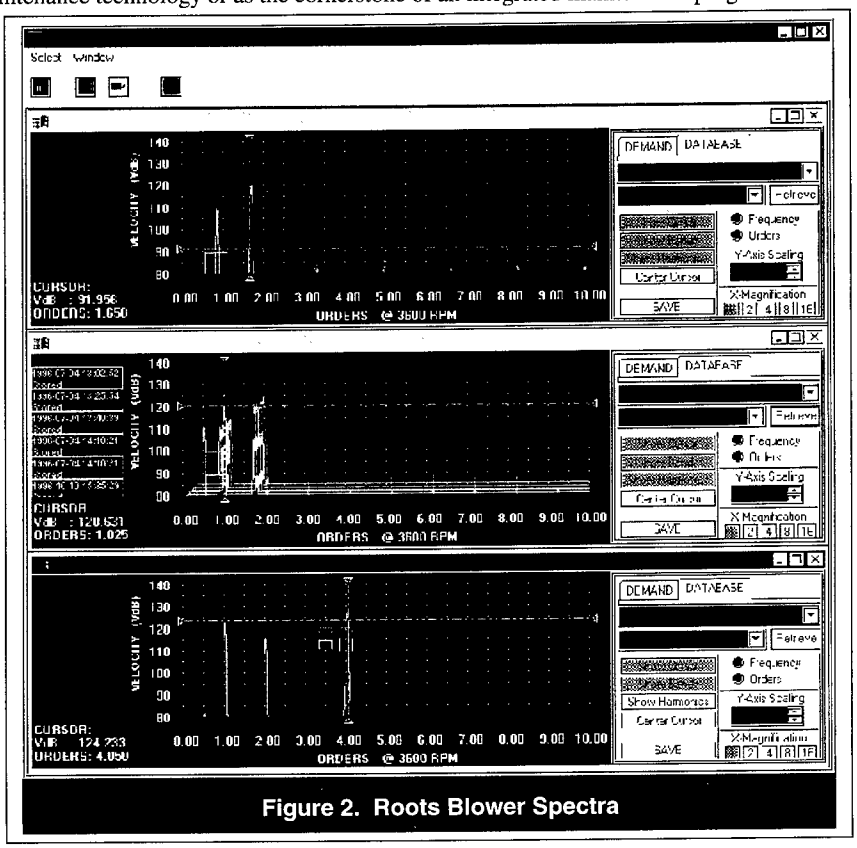


Figure 2. Roots Blower Spectra

In this case Figure 1 shows the on-line vibration approach. The values displayed are broadband levels at on the four critical bearings. RPM is a byproduct of the vibration process and so it is available for display. FFT alarms will cause the corresponding sensor cell to display in red even if the BB value is not in alarm. The FFT can then be displayed by a click on the appropriate sensor cell. See Figure 2.

This arrangement provides for online vibration monitoring and alarming. It focuses directly on the most probable sources of failure and meets the objective of protection but does not allow the tools needed to predict the conditions of surge.

Augmenting Vibration: In order to anticipate the onset of surge, several other sensors must be added. Surge is essentially an increase of outlet pressure while the inlet pressure reduces until conditions are met for reverse flow. The system then goes into oscillations of forward and reverse flow until the causes are corrected. There are several causes of surge including inadequate flow to the inlet, carry over of moisture into the blower and low inlet temperature. Figure 3 shows the addition of sensors for surge detection.

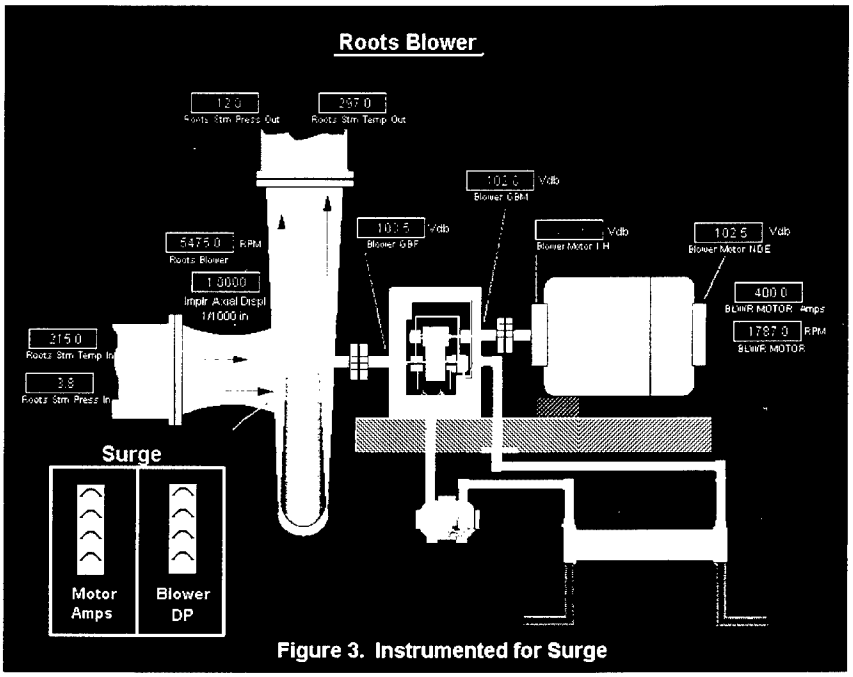


Figure 3. Instrumented for Surge

Oscillating motor amps and blower differential pressure are indicators of the surge condition and the input temperature plus differential pressure across the blower are predictors of the surge condition. With the addition of five simple sensors, the information is now available for Maintenance to understand and predict the surge

phenomenon. AS Maintenance increases its understanding of surge, they have the tools needed to enhance the expert system in an attempt to prevent it from recurring.

Involving the Operator: There is also information that is very usable to another user; the Operator. The same information used by Maintenance can (and should be) be displayed to the operator. With forewarning of the onset of surge, an Expert can be fired which instructs the operator to open the bypass or shut down the system to prevent damage. That same expert system can simultaneously alert Maintenance over the network or by modem or pager to possible problems and damage to the blower.

What has been developed is a system in which teamwork and cooperation flow naturally to achieve a common goal. In this case, the common goal is continued operation of the equipment while avoiding surge. The data is packaged for the intended user and is readily available to everyone.

Improving Performance: On a one by one basis, performance variables look at specific items. When a related group of sensors is used as a family of information in an expert system, decisions and recommendations are possible based upon subtle nuances of total system performance.

Performance of the blower is dependent on the operation of the external system (the environment) to which it supplies energy. In order to give the operator a benchmark, Motor Voltage can be added. Using the capabilities of the monitoring system, Motor Amps, Motor Voltage, and cost per kW are combined in a Virtual Sensor and displayed as Blower Cost in $/Hr to the operator. As the operator makes subtle changes to the external system, the result is displayed instantaneously as a change in the Cost in $/Hr.

Operations now has a basic tool to begin optimization of the system. Addition of other sensors, including a measure of output product can be combined in the Virtual Sensor to benchmark Cost per unit of Product.

Additional experts are also provided to guide operators and maintenance personnel when conditions exist which require corrective action. The expert system is linked into the on-line technical library and provides detailed step by step instructions (including verbal instructions and video clips) for the operator or mechanic. The expert is also linked into CMMS, for reporting, scheduling and planning. Links to DCS other online data sources are also provided.

Optimization and Protection: Maintenance can also add sensors to the blower lube oil system and gearing for further protection. Addition of these sensors will alert Maintenance to the loss of cooling efficiency, clogged filters, loss of oil pressure or flow, contamination, corrosion by-products, and out of tolerance bearing temperatures. The information is limited only to the imagination of the monitoring system users.

This integrated system presents more information than can be easily displayed so the system is split into two pages directed at specific users. Figure 1 represents the vibration viewpoint and is Maintenance oriented. Figure 4 is operator oriented but contains

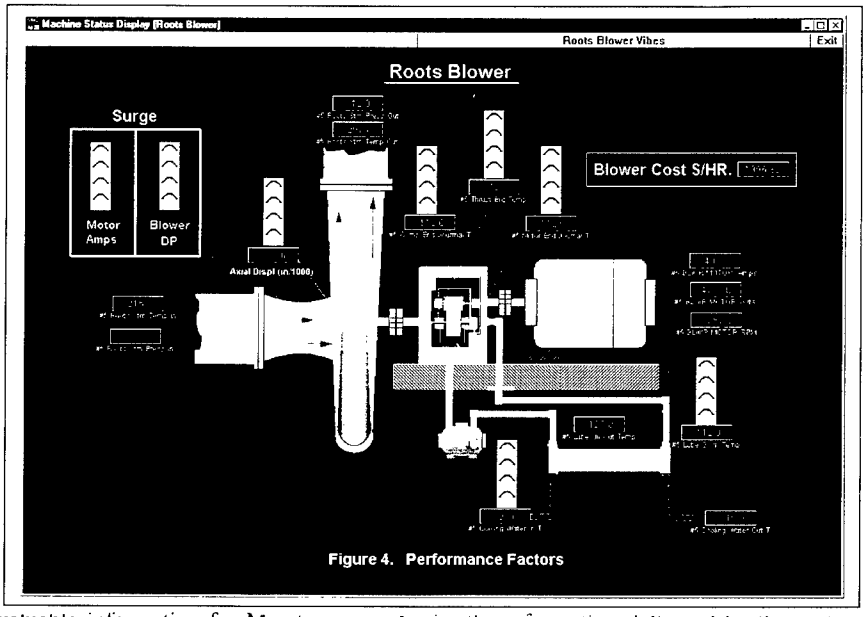


Figure 4. Performance Factors

valuable information for Maintenance. Again, the information delivered by the system facilitates a natural partnership between Operations and Maintenance.

The tools are in place now for increased system performance optimization, cause and effect studies, root cause analysis, multi-sensor trend analysis, dynamic performance alarm monitoring, and event capture. As detailed knowledge of the system is gained, it is re-invested back into the system as additional or improved experts.

Administrative and Safety Reporting Requirements: As with all projects, there are administrative and safety requirements. Reporting of required information is mandated by both outside governmental organizations and also by Company policy. Information such as cost and utilization of resources and data to support compliance with local, state and federal regulations are nearly always required. The on-line monitoring system is already gathering information and all of it is re-usable. In many cases the information may already be available in as part of a performance and maintenance monitoring scheme. In any case, additional sensors added to meet administrative requirements would be re-usable in a monitoring scheme.

It is a simple matter to add the task of gathering the required information, packaging it for the intended user and either archiving it or transmitting it to the proper agency. All of the required technologies are now available to make compliance with reporting requirements automatic and accurate. We need not burden our personnel resources with these duties.

Security and Safety Monitoring: As with Administrative and Safety reporting, monitoring of safety and security are also mandated by company policy and outside agencies. Since the technology exists to allow video and sound as valid sensor inputs, gathering the data and packaging it with other related information is a simple task.

The logic for Safety and Security are also straight forward and is easily accommodated by the Expert system. Again, there is no reason not to take advantage of the power of the monitoring system to fulfill the requirements.

Vision and Conclusion: The opportunity exists to take a program of vibration monitoring (either manual or on line) to a whole new level of sophistication and usability. The enabling technologies exist which will allow expansion of the program into areas that are limited only by the imaginations of the developers and users. Augmentation of online vibration analysis with related process data adds the ability to gain knowledge of system wide inter-relationships and also allows an expert system to make inferences from a wider knowledge base. Continual re-investment of new knowledge into the system makes it continually better.

The broad view of system operation and optimization is made available to all users of the system while the focus of vibration analysis is sharpened by knowledge of external influences. The availability of the information fosters a natural cooperation between departments to achieve common objectives.

The Vision is that information and technologies will develop to allow the formation of extended support teams made up of company personnel, equipment manufacturers, service providers and software vendors, each having access to the system information. With ever increasing communications, remote computing, and Internet capabilities, on line and real time support for the operator or mechanic from any one of the extended support team will be a click away.

References

1. "OLE for Process Control", Maintenance Technology, May,12996, pp.7-8
2. John S. Mitchell, "MIMOSA, Building the Foundation for 21st Century Optimization Asset Management", Sound And Vibration, September 1995, pp.12-13

RSAFP

FAULT DETECTION AND DIAGNOSTICS OF VIBRATING STRUCTURE - METHOD AND EXPERIMENTS

Chair: H. Günther Natke
University of Hannover, Germany

CoChair: Dirk Söffker
University of Wuppertal, Germany

MODEL-BASED DIAGNOSIS - METHODS AND EXPERIENCE

H. Günther Natke
Curt-Risch-Institute for Dynamics,
Acoustics, and Measurements, University of Hannover
Appelstr. 9 A
D-30167 Hannover, Germany

Czeslaw Cempel
Institute for Applied Mechanics
Technical University of Poznań
ul Piotrowo 3
PL-60965 Poznań, Poland

Abstract: This paper is concerned with the diagnosis of the damage and faults of mechanical systems in the diverse fields of engineering. It takes into consideration the various stages of the life cycle of a system from a diagnosis-orientated point of view and from a methodological one. Holistic modelling is the key to the methodology discussed. The holistic model includes the life behaviour (slow-time coordinate) as well as the classical dynamics (expressed with the fast-time coordinate).

Diagnosis is based on mathematical models dependent on discrete life times of the system which perform a system condition assessment in the past and in the present through an adjusted mathematical model (adaptive model), and for the future by the use of prediction with the help of adaptive models.

The principle part of the paper is the knowledge base in the form of verified and validated mathematical models which are adjusted to the states of the system at current life times. These models describe the current state of the system, permit a comparison with previous states, and therefore serve the purposes of fault detection, localization, and the cause-finding of faults and their assessment, predictions due to future forcings and trent predictions. The diagnostic decision-making based on the validated models also is mentioned. Finally, some marginal notes are made on the diagnostics of machines, and an example of a bridge illustrates the model-based diagnosis.

Key Words: Damage; diagnostics; holistic dynamics; system identification; system modifications

1 INTRODUCTION: The problem can be stated as follows: given a technical system, in particular a mechanical system, for example a machine or a structure from civil engineering, with a known (reference) condition at the life time θ_0. After a time interval $i \times \Delta\theta, i = 1, ..., N, N \epsilon \mathbf{N}$, at the lifetime $\theta_i = \theta_0 + i\Delta\theta$, a modification of the system is observed. Immediately the questions arise:

- Is the modification of the measured quantity significant?
- Where in the system (in which component, subsystem) is the modification located?
- What is the effect on the operation/serviceability of the system (assessment)?
- What causes the modification?
- Is action needed?
- What are the possible remedies?

Significant system modifications which lead to defects/faults/damage have to be detected, then diagnosis, including localization, quantification and assessment of the system modification(s), will result in finding the cause and the remedy.

In the following no distinction will be made between defects, faults and damage. They are defined here as system modifications which will be described as parameter modifications of the associated mathematical model. The system condition changes are related to the system state (variables). Some relevant and observable functions of the state variables will therefore serve as symptoms for fault detection. If the symptoms give information about the type etc. of a fault, then they are called discriminants etc.

Figure 1 gives an overview of the diagnostic process. The right branch of the figure shows the signature-based diagnostics [1] as commonly applied today. The prior knowledge of this procedure consists of the operational behaviour of the system. Here, fault localization is connected with symptom properties. The left branch of Fig. 1 is related to the model-based diagnosis. It means that a verified, validated and usable mathematical model is available. This is adjusted to every (discrete) life time of the system under observation/monitoring that is to be considered. The mathematical model which is verified and validated [2] with the use of measured data of the existing system is the optimum information available about the system. It performs the best knowledge base available for the system. Comparison of adjusted models containing possibly faulty states with a reference model will be the basis for diagnosis, cause finding and simulations of the effects of actions. In addition, it also makes possible the trend prediction of future state conditions.

An introduction to the model-based supervision and fault diagnosis of technical processes is given in [3], [4]. The methods of fault diagnosis discussed there are mainly the classification of symptoms and inference mechanisms. The model is used as the causal relation between the input and the output signals in order to calculate internal model quantities as features of fault detection. Special methods of identification are applied for process model determination.

The paper in hand summarizes the contents of the book on model-aided diagnostics by the authors [5]. It also discusses examples. Here the adjusted mathematical models are descriptions of the physical properties of the systems under investigation.

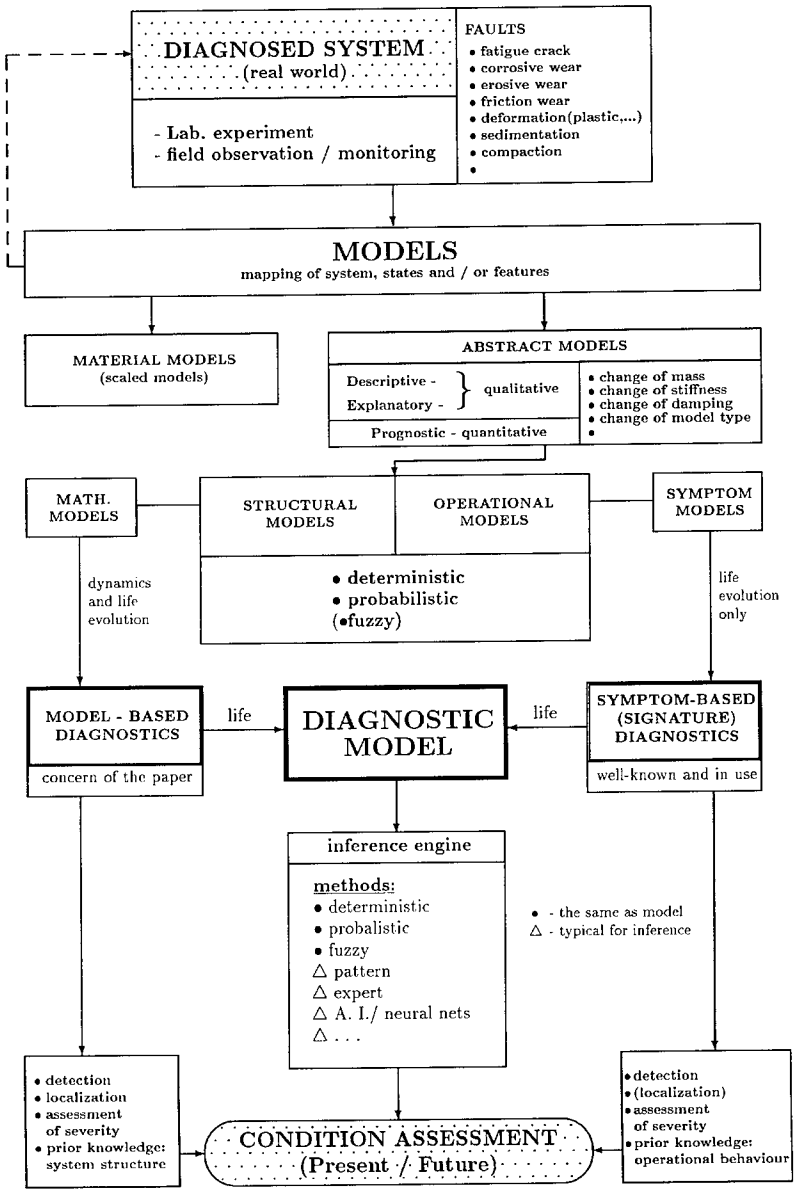


Figure 1: Overview of the diagnostic process

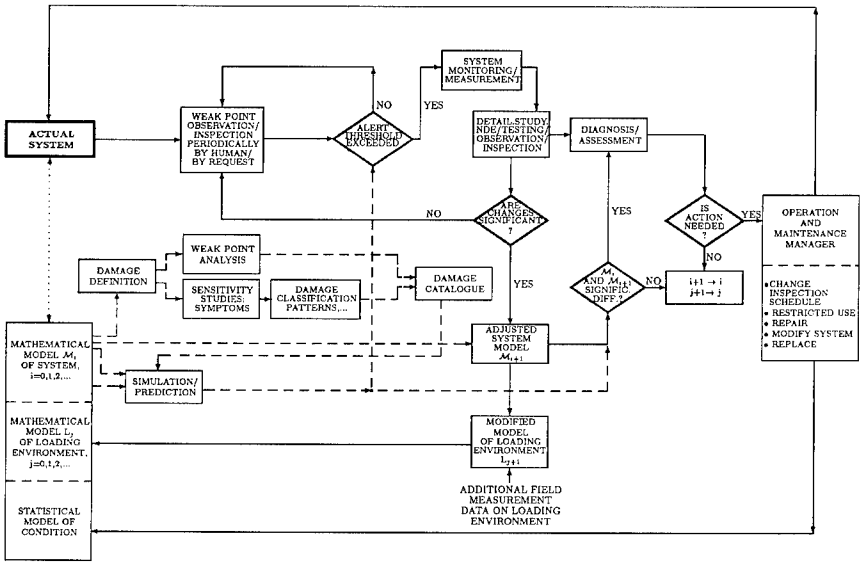


Figure 2: The model-based diagnosis

2 HOLISTIC MODELLING AND MONITORING: Figure 2 summarizes the overall model-based diagnostic procedure including prior and posteriori steps. The figure is self-explanatory. Consequently, only holistic modelling will be mentioned, and some statements will be made with regard to monitoring. For thr sake of realism, the mathematical models are restricted to spatially discretized models.

The prior mathematical model of the time-variant system is

$$M(t)\ddot{u}(t) + B(t)\dot{u}(t) + K(t)u(t) = p(t), \tag{1}$$

where $M(t), B(t)$, and $K(t)$ are life time-dependent (slowly varying with time) quadratic matrices of order n, t is the time coordinate, $u(t)$ the $(n,1)$-vector of the displacements, $p(t)$ is the $(n,1)$-vector of external forces, and dots represent the derivatives with respect to time. If the slow time coordinate θ and the fast time coordinate t are introduced (see Fig. 3), Eq.(1) can be written as

$$\mathcal{M}_i: \quad M_i\ddot{u}_i(t) + B_i\dot{u}_i(t) + K_iu_i(t) = p(t). \tag{2}$$

The index i now symbolizes the life time θ_i. The life time θ_N is generally smaller than the lifetime θ_b, the breakdown time of the system (Fig. 3).

As indicated above and as is well-known, monitoring is essential in supervision and diagnostics. Symptoms are directly and indirectly observable features with respect to faults. Symptoms are sensitive quantities with respect to the properties of the system

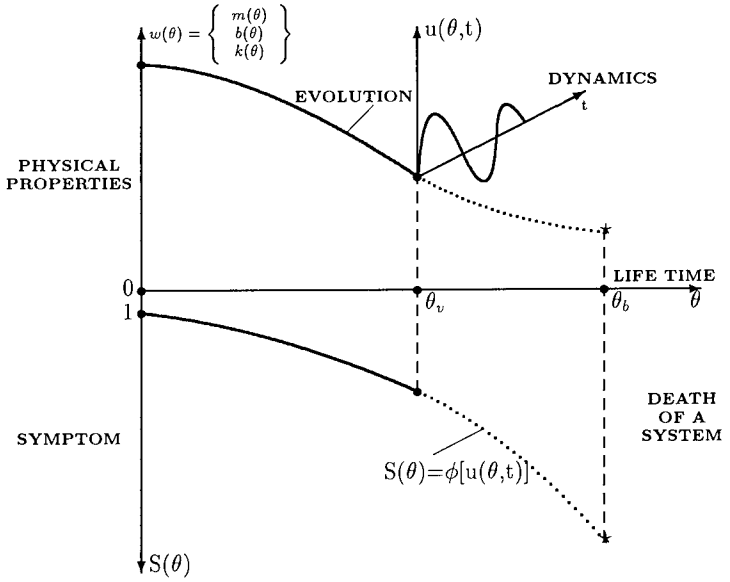


Figure 3: Slow time and fast time dynamics including the evolution of system properties and of the symptom dependent on $u(\theta, t)$

which are of interest for supervision. They are discovered by fault analysis using the prior mathematical model (see Fig. 2). Their use is shown in Fig. 2.

The related measurements can be performed continuously, periodically, or on request. The choice depends on the system properties, the goal of monitoring, and the type of damage evolution (see Fig. 3). The latter can be observed in the state modifications as abrupt, slowly changing, or sporadic. In Fig. 3 the modification is shown as slowly changing with the life time.

The system under monitoring must be appropriately equipped with suitable measuring devices. If the detected modification is significant, further investigations follow, and measurements have to be performed at least for model adjustment.

3 MODEL ADJUSTMENT TO THE CURRENT STATE AND DIAGNOSIS: Corresponding to Eq.(2), it is assumed that mathematical models are adjusted to measured states at time θ_i. This means: given the mathematical model $\mathcal{M}_{i-1}$ for the time θ_{i-1}, given measurements $\tilde{u}_i(t), \dot{\tilde{u}}_i(t), \ddot{\tilde{u}}_i(t)$, the mathematical model corrected with the given measurements $\rightarrow \mathcal{M}_i$ is sought. This is a problem of system identification called the correction, updating, calibration etc. of mathematical models. It results in parameter estimation if the irregularly varying measurement errors are modelled stochastically, and if the model structure is not changed between the time steps $i-1$ and i. Additionally, it must be assumed that, in general, fewer measured data are available than required for estimating all the elements of the pa-

rameter matrices (see Eq.(2)), and that the identification procedure should lead to fault localization.

3.1 Brief Overview of Model Correction Methods: Additionally to the references already mentioned [2], [5], the books [6] and [7] are cited in this context. In the following it is assumed that the model structure is not modified within the steps of the life time considered.

In order to draw attention to the measurements (erroneous and incomplete as mentioned before: while systematic errors are detected and corrected, irregular deviations are modelled stochastically), estimation methods are applied and subsystem modelling is introduced. The latter serves for error modelling and is a part of regularization: the introduction of a coarser parameter space which relatively increases the information available from the measurements [8]. The substructuring is done by splitting the parameter matrices in summand matrices as follows:

$$\left.\begin{array}{rcl} M & = & \sum_{\sigma=1}^{S} M_\sigma \\ B & = & \sum_{\rho=1}^{R} B_\rho \\ K & = & \sum_{\iota=1}^{I} K_\iota. \end{array}\right\} \tag{3}$$

The parametrization is now performed through the introduction of adjustment factors to every summand matrix:

$$\left.\begin{array}{rcl} M^c & = & \sum_{\sigma=1}^{S} a_{M\sigma} M_\sigma \\ B^c & = & \sum_{\rho=1}^{R} a_{B\rho} B_\rho \\ K^c & = & \sum_{\iota=1}^{I} a_{K\iota} K_\iota. \end{array}\right\} \tag{4}$$

As can be seen, the parameters chosen equal to 1 result in the prior matrices, and every estimated deviation from 1 therefore indicates a submodel modification. Consequently, this parametrization serves for localization.The summand matrices generally consist of many zero elements which can be taken into account within the computation. The parameters assembled in the vector $a = \{a_j\} = [a_{M1}, ..., a_{MS}, a_{B1}, ..., a_{KI}]^T, j = 1, ..., S+R+I$ are to be estimated: $\hat{a}$. The estimates inserted into Eq.(4) result in the following estimates of the parameter matrices: $\hat{M}, \hat{B}, \hat{K}$, which perform the adjusted mathematical model.

The estimator of the extended least squares (EWLS)[1] is recommended due to its penalty term which can be used for further regularization.

In consequence, the various adjustment procedures will result from the various residuals defined. They consist of the differences between the chosen dynamic quantities of the mathematical model to be corrected (dependent on the parameters to be estimated) and the corresponding measured values. An extended discussion of residuals will be found by the reader in [9]. Figure 4 shows the possible residuals. Modal as well as non-modal quantities can be taken. The choice depends, for example, on the type of fault and the properties of the system considered, what can be monitored

[1]This results from the Bayesian approach with normal distributions for the random variables.

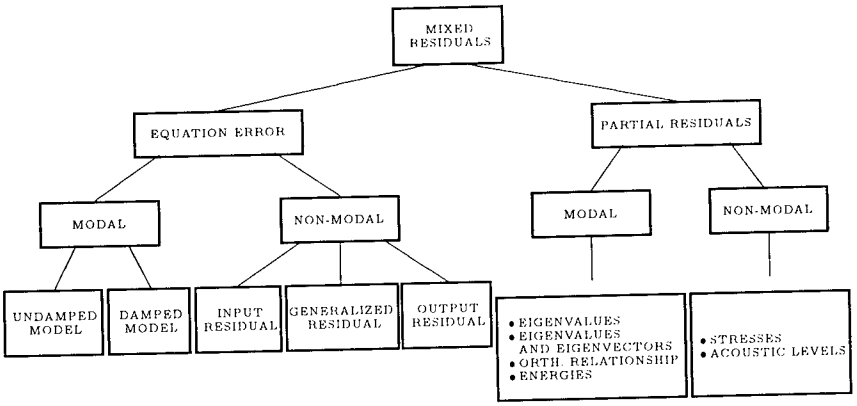


Figure 4: The various residuals

and measured etc. Of course, for significance investigations the covariances of the estimates have to be estimated. It must be said that the covariances (at least the standard deviations) can often be calculated. However, unknown systematic errors are the pitfalls of this procedure. It is also stated that the components of the residual vectors can be taken for the detection and localization of modifications. Further details can be found in the references given.

3.2 Experience with Model Correction Methods: The correction methods applied are formulated as inverse problems which, for spatially discretized models, are generally ill-conditioned. Therefore regularization methods have to be applied. Two countermeasures have already been introduced: the first is parametrization through submodels, and the second is the application of the EWLS with its penalty term for controlling the solution. The latter, of course, has to be done by additional information about the parameters. Additional information, for example, is provided by the structure and the properties of the related matrices. This stems from the sensitivity of the parameters and measurements with respect to the dynamic quantities, the measurement errors etc.

Additionally, the type of formulation of the inverse problems is of great importance. It can be done a priori well-posed or ill-posed. The flexibility formulation is thus advantageous when compared with the stiffness formulation, because if test forces are chosen, they can be chosen in that way that the related excitation matrix will be well-conditioned. Similar statements hold true for the force residual method and the response residual method (see Fig. 4).

All in all, it can be said that model adjustment with dynamic mathematical models works sufficiently well if the parametrization is adequate to the purpose, and if the measured data contain the information required for adjustment to the recent state of the system for models that are not too large in size.

3.3 Assessment and Decision-Making: First, confidence investigations with respect to the estimates have to be performed. These can be done by the aid of the Hessian matrix or the covariance matrix.

Model-based assessment and decision use simulations with the adjusted mathematical model. All the features of system analysis are available. Various loading conditions, and various faults with their effects on dynamic and static performances can be simulated. Limit stresses, pre-given acoustic levels, limit displacements etc. lead to an assessment. One has to distinguish between the properties of dynamic and static models. Dynamic models are global, energy-equivalent models, generally with a few degrees of freedom (number of modes), while static models with many (static) degrees of freedom allow the representation of local properties.

Cause finding and trend prediction are also based on the adjusted models. Trend prediction can be performed deterministically by extrapolation of the $\mathcal{M}_i$, or statistically. Decision-making follows the classical rules or can be based on fuzzy logic which is discussed in [5] (but not in great details).

4 APPLICATIONS TO MECHANICAL AND CIVIL ENGINEERING: First, an academic example in [5] is discussed. It is a finite element (FE) model of a satellite. The damage is introduced by parameter modifications of the prior mathematical model. The measurements are also simulated, and include stochastical errors. The localization of the damage and the adjustment of the mathematical model to the simulated measured data are discussed. The application and the effects of regularization on the estimates are shown, taking data sensitivities into account.

4.1 Mechanical Engineering: Some differences exist between diagnostics and condition monitoring in mechanical and civil engineering. One difference lies in the fact that, in addition to the structural parameters, operational parameters like rotation speed and pressure play an important role for systems of mechanical engineering. Some prior information is available for machines. For example, the detection of shaft misalignment can be done by observing the shaft's relative vibration with the probability 0.8, but when observed at the bearing support the probability of detection will be as low as 0.1. The same applies to machine casing. With respect to the direction of vibration observation for this fault, the axial direction has the probability 0.5, the horizontal one 0.3, and the vertical one 0.2.

Based on these data one can state that for machine diagnostics the fault location problem is currently not as urgent as for civil engineering systems (with the exception of special cases of turbosets, for example). This may also be due to the much smaller dimensional scale of machines in comparison with buildings (approximately ten times smaller). However, this does not mean that model-supported diagnostics is not needed for machines and equipment. Apart from this, by having a good holistic model of some critical machines we can verify and validate the available historical data on machine behaviour gathered previously. With the same model we can simulate the condition degradation during the modelled usage, and in this way we can build and then validate the models, and assess the machine conditions as summarized below:

- finding weak structural points where early degradation occurs, so that they can be modified in successive redesign. If this is not the case, they will be the first places for the installation of condition monitoring equipment
- finding and choosing the best damage-orientated symptom S, (for detection, location and assessment), by simulating a number of possible measurable quantities, and by processing their signals accordingly
- choosing the best symptom operator $\phi(.)$ for the given case of wear, together with its shape factor γ, which allows us to adjust the evolution of mass, spring, damping coefficients in our holistic model. This will also serve for the prognosis of symptom behaviour in the given operational condition
- validating by response simulations the fault detection/assessment decisions obtained from previous statistical data.

Some of the enumerated tasks and cases are illustrated below.

The model-based diagnostics for rotor machinery including turbosets will be discussed briefly.

Rotor machinery: Diagnostics is always understood as 'detection - localization - assessment'. Localization is not so important due to the statistics already gathered, and also to the much smaller physical dimensions compared with other systems, which provides another range of dynamic phenomena and wave behaviour in particular. Hence, with respect to the model-based condition monitoring of machines only the detection and the assessment problem have been studied extensively so far. As one of the latest examples of such research, reference is made briefly to the main idea and findings of paper [10]. This paper, with an extended literature survey, concerns the investigation of a simple supported rotating shaft loaded by a concentrated mass (ring), which is the model of the rotor. The problem is to develop and/or improve some known methods of signal processing for the task of the detection and assessment of rotor unbalance and shaft crack. It was shown by analytical and experimental investigations that the extended Kalman filter and the modified instrumental variable method can be used succesfully for the detection and the assessment of both faults with acceptable accuracy when compared with the real condition introduced experimentally. The location problem is not investigated.

Turbosets: The oldest application of condition monitoring and diagnostics in engineering began in power stations, and with steam turbines in particular. Initially (up to the Sixties), this produced the so-called Turbine Supervisory Equipment (TSE) based on transducers, analogue signal conditioning and processing. They mainly gave some alarm signals and sometimes caused a preselected emergency shut-off when the symptom limit value was exceeded. But the main work of the diagnostic interpretation of elaborated symptoms was carried out by diagnostic staff.

The rapid development in electronics and computer signal processing has transformed the TSE of turbosets into Turbine Monitoring Equipment (TME), with integrated tailor-made computers for signal processing especially devoted to turbosets (shaft trajectories, cascade spectrum of run-downs and run-ups, symptom alarm and limit

values, etc). Since the Eighties this has been the standard equipment of most modern power stations, and due to the long application of both systems in practice many statistical data from turbine diagnostics have been gathered and elaborated. This currently makes possible the application of knowledge engineering and expert systems, in particular, to the diagnostics of turbosets. Due to the progress in identification theory and practice, and in computer-based systems of the FE-method, and to the successful modelling of some interaction effects typical to slide bearing machinery, it was possible to elaborate computer models for the simulation of nearly all dynamic phenomena important in diagnostics and in the supervision of operating turbosets. Another obstacle to the modelling of rotor machinery with journal bearings should be mentioned here: the highly nonlinear effects in journal bearings and shaft sleeves which were elaborated successfully quite recently.

Hence, the possibility of elaborating some type of model-supported diagnostics of turbosets is now emerging. It is tentatively called the 'Intelligent Diagnostic System' which, together with the expert system, incorporates and integrates two knowledge bases: it includes a validated dynamic model of the turboset, which can be used for checking the statements (assessment) previously made, for confirming the recent state conditions (symptom values) and for prediction.

With regard to the current state of the development of dynamic models of turbosets in general, it holds true for the:

- simulation of some dynamic phenomena (vibration, support reaction, temperature, etc.) in good agreement with the experiment
- modelling and simulation of some abnormal effects of operation, i.e. faults
- description of the resulting simulated condition in terms of a set of symptoms, which already have a good diagnostic meaning.

In particular, the dynamic models permit one to simulate such faults as:

- unbalance at any point of the shaft and rotor
- different types of oil instabilities in journal bearings
- the evolution of the operational line of the shaft and the resulting position of bearing supports
- cracks in a rotor and shaft and the evolution of the cross-section due to erosion
- the change in the stiffness of bearing supports
- abnormal electromagnetic forcings from the side of the generator
- condition evolution in journal bearings
- the misalignment of the shaft and couplings.

This can be checked under the given simulated operational conditions, and will validate the existence and intensity of faults in terms of diagnostic symptoms and operational parameters. The dynamic model also permits one to simulate the dynamic behaviour under measured operational conditions and to look for the causes of observed abnormalities. In addition, simulations with modified operational parameters dependent on the structural parameters lead to an assessment of future conditions.

Table I: Comparison of identified and calculated eigenfrequencies and modes

DESCRIPTION OF MODES	EIGENFREQUENCY [Hz]		DIFFERENCE %	MAC
	CALCULATED	MEASURED		
1.VERTICAL BENDING	0.79	0.756 ± 0.0021	3.3	0.997
2.VERTICAL BENDING	1.15	1.157 ± 0.0033	0.6	0.907
3.VERTICAL BENDING	1.32	1.307 ± 0.0051	1.3	0.887
4.VERTICAL BENDING	1.99	1.911 ± 0.0043	4.1	0.932
1.TORSION	1.21	1.173 ± 0.0054	3.2	0.996
2.TORSION	1.56	1.574 ± 0.0065	0.9	0.972
3.TORSION	2.35	2.274 ± 0.0069	3.3	0.832
4.TORSION	2.56	2.500 ± 0.0053	2.4	0.887

The computational expenditure is very high. For low-pressure casing the dynamic FE-model has 60,000 DOFs, and it contains at least two nonlinear and non-stationary models of slide bearings.

Consequently, model-based diagnosis in its general sense can be a very powerful tool for application in mechanical engineering systems, and this field is also still at the research stage. There are reports that several turbine producers are working extensively on the appropriate developments.

4.2 Civil Engineering: The application here concerns an existing cable-stayed bridge (Norderelb bridge) which was subjected to an eigenvibration analysis and a modal test. A detailed description of the development of an expert system can be found in [11].

Comparison of Calculated and Estimated Eigenquantities: Table I contains the calculated and identified results. The mode description characterizes the main deflections of the mode. The relative differences between the calculated and identified eigenfrequencies are relatively small. The last column contains the values of the MAC[2] for the calculated and measured eigenvectors. The MAC is the cosine of the angle between the calculated and identified eigenvectors. In the ideal case this value is identical to 1. As can be seen, the eigenvectors agree relatively well with each other.

Dependent on the catalogue of faults, it now has to be decided whether the FE-model used is in fact usable. If not, it has to be adjusted in order to serve as a reference model. Correction of the FE-model will not be performed here, because, as will be seen next, the FE-model is usable for the detection and localization.

Simulated Damage and its Description: The lower chord of the main girders is build as a bottom flange. The cross-section is closed in the middle field of the bridge, and in this area the stiffness is mainly determined by the ground plate. However,

[2]Modal assurance criterion.

Table II: Comparison of the eigenquantities of the damaged system and the reference values

DESCRIPTION OF MODES	EIGENFREQUENCY [Hz]		DIFFERENCE %	MAC
	REFERENCE	DAMAGED SYSTEM (SIMULATED)		
1.VERTICAL BENDING	0.765	0.79	3.3	0.985
2.VERTICAL BENDING	1.157	0.89	-23.0	0.860
3.VERTICAL BENDING	1.307	1.23	-5.9	0.562
4.VERTICAL BENDING	1.911	2.04	6.8	≈ 0
1.TORSION	1.173	1.23	4.9	0.994
2.TORSION	1.574	1.68	6.7	0.960
3.TORSION	2.274	2.37	4.2	0.818
4.TORSION	2.500		-5.2	0.120
5.TORSION	2.652	2.60	-2.0	0.938
6.TORSION	3.600	3.50	-2.8	0.922

in the end field, defined by the large deflection of the 2nd vertical bending mode, the bottom flanges of the longitudinal girders are essential. It is assumed that these flanges are completely missing: the stiffnesses of the bottom flanges in the end field are removed in the related FE-model.

The damage simulated in this way leads to the simulated eigenquantities as shown in Table II. Again, a comparison of the eigenquantities of the (simulated) damaged system with the corresponding quantities of the reference model is now presented. The effect is enormous. The 2nd bending eigenfrequency shows a reduction of 23% as an effect only of the stiffness reduction in the end field. The related mode does not differ in its shape from the corresponding mode of the undamaged system. Most of the MAC values differ significantly from 1. This is mainly due to the changed signs of the modal vector components (not shown). By comparison with the reference model, the number of DOFs (of normal modes) is now reduced. This is astonishing when one looks only at the stiffness modification in the end field. Here a sophisticated decision based on a significance test is unnecessary. This result shows the strong stiffness coupling of the various fields of the bridge.

The eigenquantities permit the detection and localization of the assumed damage, and therefore they are good symptoms in the example considered. The 2nd bending mode shows the main deflection in the end field (damaged as well as undamaged). The eigenfrequency of the damaged system is much smaller than that of the undamaged system; consequently, something has happened regarding stiffness reduction (or added masses) in this vibrating field. Inspection will confirm the result.

This means that the reverse method, namely the adjustment of the reference model, can easily be performed by looking at the used and known FE-models, and, of course,

when taking into account the modification of the number of degrees of freedom. The diagnosis is easy to perform with the available knowledge. Assessment will lead to action.

5 CONCLUSIONS: Signal and signature supported diagnosis are well known and have already been introduced for the monitoring and diagnosis of (vibrating) systems in operation. Higher demands with respect to the safety and lifetimes of systems require efficient and automatically elapsing tools for assessment and decision. From the economic point of view, the costs of development, and the costs during operation/service of a system, can be reduced if the faults of a system are detected early and their evolution is known. Assessment of such faults then leads to action in time, and avoids subsequent costs. The costs of monitoring itself can also be reduced if the monitoring is performed state-dependent and event-dependent, and not periodically. With regard to machines the importance of monitoring and diagnostics is obvious, because condition monitoring was introduced in this field of application much earlier than in other technical fields. For civil engineering systems it is obvious that the deteriorating infrastructure and environment both pose challenging problems for engineering and diagnostics as well. It would be a great contribution to society if engineers could save even a fraction of the percentage of the cost by improving the present method of maintaining the infrastructure and by designing in a new, more service-orientated manner. Efficient diagnosis procedures are the basis for economic renewal engineering.

Model-based diagnostics is theoretically an optimum method for damage detection, localization and assessment, because verified, validated and usable mathematical models at every state condition are the best knowledge base available. The expenditure is great. However, dependent on the complexity and criticality of the system and its societal and economic importance, the expenditure on monitoring, measurements and the subsequent computations can be much lower than any other approach required to achieve the necessary safety and performance, not to mention the breakdown costs.

The recent advancement in computational engineering (software, hardware, and numerical methods), the development in measuring techniques and in system identification encouraged the authors to propose and discuss the system identification-based methodology for diagnosis.

As already stated, model-based diagnosis is discussed from a methodological point of view; this means that the intention is to provide the reader with a stimulus and an overview of some methods, so that he/she will be able to start the investigation of his/her particular problem.

Modified forces and environmental conditions have to be taken into account, of course. Environmental parameters and (resulting) special conditions, such as pre-stressing, have to be known and registered or taken as additional loads in the particular case.

The quality of the models used is decisive in the procedures discussed. The results will only be as informative as the models are able to describe the required properties.

One has to distinguish between global identification and local identification. This is discussed through the distinction between dynamic models and static models when subsystem modelling is considered. Dependent on the criteria decisive for assessment, it can be necessary, after fault localization, to use an expanded static submodel for that part of the system where the damage affects safety, performance, or comfort. Or it may be necessary, too, to introduce additionally a damage model in order to obtain detailed information about the extent of the damage. These facts require adaptive models, but it may be a long path until a procedure which uses such adaptive models is achieved.

Adaptive models should take into account expected and predicted damage and the resulting changed external forces in order automatically to perform an assessment and trend prediction.

The content of the book [5] is restricted to linear models, although nonlinear system behaviour is partially discussed. In each application a check has to be made as to whether this assumption will be approximately valid (property-dependent and purpose-equivalent). For systems behaving nonlinearly, the reader's attention is drawn to [12] and the references cited in it.

6 REFERENCES:

1. Ben-Haim, Y., Cempel, C., Natke, H.G., Yao, J.T.P., 1993, Evaluation of diagnostic methods; in [12], 1-11
2. Natke, H.G., 1996, Substructuring within model verification and validation; Proc. of the Intern. Congress "New Advances in Modal Synthesis of Large Structures. Nonlinear, Damped and Nondeterministic Cases." Ed. L. Jezequel, Balkema Rotterdam, Brookfield
3. Isermann, R., 1996, Modellgestützte Überwachung und Fehlerdiagnose technischer Systeme; atp - Automatisierungstechnische Praxis 38, 5 und 6, 9-20 und 48-57
4. Cempel, C., 1991, *Vibroacoustic Condition Monitoring*, Ellis Horwood Chicester, UK
5. Natke, H.G., Cempel, C., 1996, *Model-Aided Diagnosis of Mechanical Systems - Fundamentals, Detection, Localization, Assessment*, Springer-Verlag
6. Friswell, M.I., Mottershead, J.E., 1995, *Finite Element Model Updating in Structural Dynamics*, Kluwer Academic Dordrecht
7. Natke, H.G., 1992, *Einführung in Theorie und Praxis der Zeitreihen- und Modalanalyse - Identifikation schwingungsfähiger elastomechanischer Systeme*; Friedr. Vieweg & Sohn Braunschweig, Wiesbaden
8. Natke, H.G., 1993, On regularization methods within system identification; in: M. Tanaka, H.D. Bui (eds.), *Inverse Problems in Engineering Mechanics*; Springer-Verlag Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest, 3-20
9. Natke, H.G., Lallement, G., Cottin, N., Prells, U., 1995, Properties of various residuals within updating of mathematical models; Inverse Problems in Engineering, Vol. 1, 329-348

10. Seibold, S., 1995, *Ein Beitrag zur modellgestützten Schadensdiagnose bei rotierenden Maschinen*, Fortschrittsberichte VDI, Reihe 11: Schwingungstechnik, Nr. 219, VDI-Verlag Düsseldorf
11. Natke, H.G., Doll, H., Hildebrandt, P., Schütze,M., 1995, Bridge condition assessment using an expert system; Structural Engineering Review, Vol. 7, No. 3, 165-180
12. H.G. Natke, G.R. Tomlinson, J.T.P. Yao (eds.), 1993, *Safety Evaluation Based on Identification Approaches Related to Time-Variant and Nonlinear Structures*, Vieweg Verlag Braunschweig, Wiesbaden

DAMAGE DETECTION OF DYNAMICAL SYSTEMS BY SVD AND INFLUENCE COEFFICIENTS

A. Lenzen
Ruhr University Bochum IA/6/47
D-44780 Bochum, Germany

Abstract : This paper deals with a new method to investigate dynamical systems based on singular value decomposition and influence coefficients, which are deduced from the integral solution of the state space equation. The investigation can be used for a damage detection of the dynamical system.

Key Words : Damage detection; structural dynamics; system identifikation; singular value decomposition

INTRODUCTION : Most buildings of structural and mechanical engineering are dynamical constructions which can be analyzed by experimental measurement to registrate structural changes. Constructions are damaged as results of overloading, fatigue and aging. With measuring vibrations and methods of system identification it is possible to determine and localize these damages to repair the buildings.
In opposite to numerical models with methods of finite elements, the new method to investigate dynamical systems and registration of structural changes is based on the black box model or state space model as known in system theory. The physical system, a real dynamic construction, is measured and described by parameters of the state space realization. The parameters of the numerical model are estimated by the generalized singular value decomposition out of input and output quantities of the black box model. The parameters of the state space realization are not unique and allow no localization of structural changes.
The introduced influence coefficients, evaluated from the integral equation of the model, allow a damage detection of the dynamical system under investigation. The suitability of the method is demonstrated by an experimental model.

SINGULAR VALUE DECOMPOSITION : The singular value decomposition (SVD) is a numerical algorithm [1], which enables the decomposition of a matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ into two unitary matrices of complex numbers (two orthogonal matrices of real numbers) and one diagonal matrix. The first stable computer program was presented in 1969 by Businger and Golub.
The theoretical generalization of the SVD of two matrices $\mathbf{A} \in \mathbb{K}^{m \times n}, \mathbf{B} \in \mathbb{K}^{p \times n}$, referred to as GSVD (Generalized Singular Value Decomposition) was first described by Paige and van Loan [7] [6] in the years 1976 and 1981.
For a matrix

$$\mathbf{A} \in \mathbb{K}^{m \times n} \tag{1}$$

the decomposition for the SVD results in two unitary matrices

$$\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_m] \in \mathbb{K}^{m \times m} \text{ and } \mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_n] \in \mathbb{K}^{n \times n}, \tag{2}$$

so that for $p = \min(m, n)$ the equation

$$\mathbf{U}^* \mathbf{A} \mathbf{V} = \operatorname{diag}(\sigma_1, \cdots, \sigma_p) \in \mathbb{R}^{m \times n}, \tag{3}$$

is valid whereby $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.
The σ_i are referred to as singular values, the $\mathbf{u}_i$ as left singular vectors and the $\mathbf{v}_i$ as right singular vectors. The rank of a matrix can be determined by using the SVD. If the matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$ occupies the rank r then with the aid of the unitary matrices $\mathbf{U} \in \mathbb{K}^{m\times n}$, $\mathbf{V} \in \mathbb{K}^{n\times n}$ and the diagonal matrix $\mathbf{\Sigma}_A \in \mathbb{R}^{m\times n}$ it is possible to reduce the matrix $\mathbf{A}$ into factors

$$\mathbf{\Sigma}_A = \begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & \ddots & \\ 0 & & & & 0 \end{bmatrix} = \mathbf{U}^*\mathbf{A}\mathbf{V} \tag{4}$$

whereby the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.
The quotient of two singular matrices $\mathbf{A}\mathbf{B}^{-1}$ can be presented as a generalized SVD. Two matrices

$$\mathbf{A} \in \mathbb{K}^{m\times n} \tag{5}$$

and

$$\mathbf{B} \in \mathbb{K}^{p\times n}, \tag{6}$$

with the same number of rows can be separated into two unitary matrices

$$\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_m] \in \mathbb{K}^{m\times m} \text{ and } \mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_p] \in \mathbb{K}^{p\times p}, \tag{7}$$

and one invertible matrix

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{K}^{n\times n} \tag{8}$$

so that

$$\begin{aligned} \mathbf{U}^*\mathbf{A}\mathbf{X} &= \operatorname{diag}(\alpha_1, \cdots, \alpha_r, \cdots \alpha_{r+s}, \cdots, \alpha_k, 0) \quad \in \mathbb{R}^{m\times n} \\ \mathbf{V}^*\mathbf{B}\mathbf{X} &= \operatorname{diag}(\beta_1, \cdots, \beta_r, \cdots, \beta_{r+s}, \cdots, \beta_k, 0) \quad \in \mathbb{R}^{p\times n} \end{aligned} \tag{9}$$

and further that in the case of nontrivial pairs of singular values the following can be said

$$\begin{aligned} &\alpha_i = 1 \quad , \quad \beta_i = 0 \qquad i = 1, \ldots, r, \\ &\alpha_i < 1 \quad , \quad \beta_i > 0 \qquad i = r+1, \ldots, r+s, \\ &\alpha_i = 0 \quad , \quad \beta_i = 1 \qquad i = r+s+1, \ldots, k. \end{aligned} \tag{10}$$

Further forms of SVD algorithms can be found in the newly literature of numerical analysis.

DYNAMICAL SYSTEMS : Dynamic constructions in mechanical and structural engineering are treated numerically by discretising them by methods of finite elements. The mathematical model describes the reality, under assumption of the theory of elasticity with a differential equation of second order.
The equation of motion for a finite element model, if elastic material laws and small deformations are presumed, can be found as

$$\mathbf{M}\ddot{\mathbf{y}}_{(t)} + \mathbf{D}\dot{\mathbf{y}}_{(t)} + \mathbf{K}\mathbf{y}_{(t)} = \mathbf{u}_{(t)} \tag{11}$$

with $\mathbf{M}$: mass matrix, $\mathbf{D}$: damping matrix, $\mathbf{K}$: stiffness matrix, $\mathbf{y}, \dot{\mathbf{y}}, \ddot{\mathbf{y}}$: displacement, velocity, acceleration and $\mathbf{u}$: force vectors. The dimensions of the matrices correspond to the degrees of freedom of the chosen models $\mathbf{y}, \mathbf{u} \in \mathbb{R}^s$; $\mathbf{M}, \mathbf{D}, \mathbf{K} \in \mathbb{R}^{s \times s}$.
It is possible to transfer a differential equation of second order into a differential equation of first order, thus gaining access to the state space realization.

$$\begin{bmatrix} \dot{\mathbf{y}}_{(t)} \\ \ddot{\mathbf{y}}_{(t)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{(t)} \\ \dot{\mathbf{y}}_{(t)} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{M}^{-1} \end{bmatrix} \mathbf{u}_{(t)} \tag{12}$$

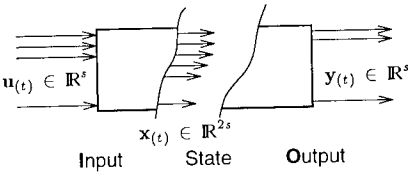
$$\dot{\mathbf{x}}_{(t)} = \mathbf{A}\,\mathbf{x}_{(t)} + \mathbf{B}\,\mathbf{u}_{(t)}$$

with

$$\mathbf{y}_{(t)} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \mathbf{x}_{(t)}, \qquad \mathbf{y}_{(t)} = \mathbf{C}\mathbf{x}_{(t)} \quad . \tag{13}$$

As the state space [2] [8] of this model can be interpreted in direct physical terms of displacement and velocity thus it must be considered as a special case of a general representation of a vector differential equation with $\mathbf{A} \in \mathbb{R}^{2s \times 2s}$, $\mathbf{B} \in \mathbb{R}^{2s \times s}$, $\mathbf{C} \in \mathbb{R}^{s \times 2s}$, $\mathbf{x} \in \mathbb{R}^{2s}$.
The parameters of the equation are known factors derived by the laws of mechanics thus this state space equation must be seen as a synthetic mathematical model, figure 1, representing reality.

unknown :
$\mathbf{A} \in \mathbb{R}^{n \times n}$
$\mathbf{B} \in \mathbb{R}^{n \times q}$
$\mathbf{C} \in \mathbb{R}^{p \times n}$
known :
I/O p, q



known :
$\mathbf{A} \in \mathbb{R}^{2s \times 2s}$
$\mathbf{B} \in \mathbb{R}^{2s \times s}$
$\mathbf{C} \in \mathbb{R}^{s \times 2s}$
unknown :
I/O

Figure 1: Analysis (left) / Synthesis (right)

The state space, in contrast to the above example, can generally be regarded as a representation which does not lend itself to a physical interpretation. In this case the dimension n of the state space is an unknown. The state space model in its entirety of $\mathcal{S}\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ is however depicted in reality as a mathematical model according to the principles of cause and effect. Input and output quantities $\mathbf{y}_{(t)}$, $\mathbf{u}_{(t)}$ connect the state space to the "outside world" through the parameter matrices $\mathbf{B}$ and $\mathbf{C}$ by means of the following two equations:

$$\dot{\mathbf{x}}_{(t)} = \mathbf{A}\mathbf{x}_{(t)} + \mathbf{B}\mathbf{u}_{(t)} \tag{14}$$

$$\mathbf{y}_{(t)} = \mathbf{C}\mathbf{x}_{(t)}. \tag{15}$$

If the parameters $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times q}$, $\mathbf{C} \in \mathbb{R}^{p \times n}$ are unknown, then this method of state space realization is also referred to as the black box model. By determining the unknown parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ with the input and output quantities it is possible to analyze a real system. This is the reason why the black box method can be regarded as an analytical mathematical model, figure 1. The state space of the black box model is not a unique description and it cannot be directly interpreted in terms of general physics, because the input

and output is measured and different state space models may have the same result. It is always possible to transform the parameters of the models $\mathcal{S}\{\mathbf{A},\mathbf{B},\mathbf{C}\}$ with the coordinate transformation

$$\mathbf{x}_{(t)} = \mathbf{P}_{(t)}\boldsymbol{\xi}_{(t)}, \tag{16}$$

into a alternative state space realization $\mathcal{S}_2\{\mathbf{P}_{(t)}^{-1}\mathbf{A}\mathbf{P}_{(t)} - \mathbf{P}_{(t)}^{-1}\dot{\mathbf{P}}_{(t)}, \mathbf{P}_{(t)}^{-1}\mathbf{B}, \mathbf{C}\mathbf{P}_{(t)}\}$. The solution to the vector differential equation of first order is known, in the case of linear time invariant systems it is

$$\begin{aligned}
\mathbf{y}_{(t)} &= \boldsymbol{\mathcal{P}}_{(t)}\mathbf{x}_0 \qquad + \int_{t_0}^{t} \mathbf{T}_{(t,\sigma)}\ \mathbf{u}_{(\sigma)}\,\mathrm{d}\sigma \\
&= \boldsymbol{\mathcal{P}}_{(t)}\mathbf{x}_0 \quad + \boldsymbol{\mathcal{P}}_{(t)} \int_{t_0}^{t} \boldsymbol{\mathcal{Q}}_{(\sigma)} \quad \mathbf{u}_{(\sigma)}\,\mathrm{d}\sigma \\
&= \boldsymbol{\mathcal{P}}_{(t)}\mathbf{x}_0 + \mathbf{C}e^{\mathbf{A}(t-t_0)} \int_{t_0}^{t} e^{\mathbf{A}(t_0-\sigma)}\mathbf{B}\mathbf{u}_{(\sigma)}\,\mathrm{d}\sigma
\end{aligned} \tag{17}$$

The hankel transformation represents a special case of state space realization, figure 2. At point t_0 the past state is transformed into the future state.

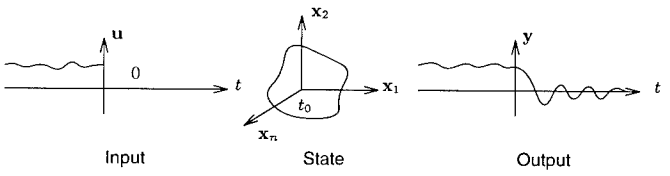


Figure 2: Hankel transformation

$$\mathbf{y}_{(t)} = \boldsymbol{\mathcal{P}}_{(t)}\mathbf{x}_0 \quad ; \quad \mathbf{x}_0 = \int_{-\infty}^{t_0} \boldsymbol{\mathcal{Q}}_{(\sigma)}\mathbf{u}_{(\sigma)}\,\mathrm{d}\sigma \tag{18}$$

The special feature is that the input $\mathbf{u}_{(t)}$ is turned off at point t_0, thus it can be said in general terms that the hankel transformation can be regarded as a generalized impulse response, which forms the subset to this.

The input $\mathbf{u}_{(t)} \in \mathbb{R}^q$ is excited in succession through special signals, for example a impulse function. Then the time synchronous measured output $\mathbf{y}_{(t)} \in \mathbb{R}^p$ will represent the information of the dynamical system in form of the hankel transformation.

$$\mathbf{x}_{0i} = \int_{-\infty}^{t_0} e^{\mathbf{A}(t_0-\sigma)}\mathbf{B} \begin{bmatrix} 0 \\ e_i = 1 \\ 0 \end{bmatrix} \delta_{(t_0-\sigma)}\,\mathrm{d}\sigma = \mathbf{b}_i \quad , \quad i = 1,\ldots,q \tag{19}$$

The state space vector $\mathbf{x}_0 \in \mathbb{K}^n$ transmits the information of the mathematical model from the input to the output.

$$\mathbf{y}_{(t),i} = \mathbf{C}e^{\mathbf{A}(t-t_0)}\mathbf{x}_{0i} = \mathbf{C}e^{\mathbf{A}(t-t_0)}\,\mathbf{b}_i \quad , \quad i = 1,\ldots,q \quad . \tag{20}$$

After assembling the single measured output vectors $\mathbf{y}_{(t),i} \in \mathbb{R}^p$ from the selected input channels $\mathbf{u}_{(t),i} \in \mathbb{R}^q$; $i = 1,\ldots,q$ they will serve as information to build the dynamic model.

$$\left[\ \mathbf{y}_{(t),1} \,\middle|\, \mathbf{y}_{(t),2} \,\middle|\, \cdots \,\middle|\, \mathbf{y}_{(t),q}\ \right] = \mathbf{Y}_{(t)} = \mathbf{C}e^{\mathbf{A}(t-t_0)}\,\mathbf{B} \in \mathbb{R}^{p\times q} \tag{21}$$

This formulation is corresponding to the parameters of the state space model. As a matter of fact the continuous signals $\mathbf{Y}_{(t)}$ are sampled by a digital computer. The equidistant discrete measurements, with $t_{j-1} < t_j < t_{j+1}$ and $\Delta t = t_{j+1} - t_j$, $j = 0,\ldots,\infty$, can be formulated as followed

$$\begin{aligned}
\mathbf{Y}_0 &= \mathbf{Y}_{(t_0)} = \mathbf{C}\,e^{\mathbf{A}(t_0-t_0)}\,\mathbf{B} \\
\mathbf{Y}_1 &= \mathbf{Y}_{(t_1)} = \mathbf{C}\,e^{\mathbf{A}(t_1-t_0)}\,\mathbf{B} = \mathbf{C}\,e^{\mathbf{A}\Delta t}\,\mathbf{B} \\
&\ \ \vdots \\
\mathbf{Y}_j &= \mathbf{Y}_{(t_j)} = \mathbf{C}\,e^{\mathbf{A}(t_j-t_0)}\,\mathbf{B} = \mathbf{C}\,(e^{\mathbf{A}\Delta t})^j\,\mathbf{B}.
\end{aligned}$$

The equations illustrate the interrelation between the measurement and the model parameters. If the system under consideration is time invariant a so called hankel matrix can be constructed out of the sampled output data.

$$\mathbf{H} = \left[\begin{array}{c|cccc} \mathbf{Y}_0 & \mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_\infty \\ \hline \mathbf{Y}_1 & \mathbf{Y}_2 & & & \\ \mathbf{Y}_2 & & & & \\ \vdots & & & & \\ \mathbf{Y}_\infty & & & & \end{array}\right] \tag{22}$$

The theoretical dimensions of the hankel matrix are infinite.
In opposite to the infinite measured hankel matrix the numerical formulation of the mathematical model lends to a block-dyadic structure.

$$\mathbf{H} = \left[\begin{array}{c} \mathbf{C}(e^{\mathbf{A}\Delta t})^0 \\ \hline \mathbf{C}(e^{\mathbf{A}\Delta t})^1 \\ \hline \mathbf{C}(e^{\mathbf{A}\Delta t})^2 \\ \hline \vdots \\ \hline \mathbf{C}(e^{\mathbf{A}\Delta t})^\infty \end{array}\right] \left[\ (e^{\mathbf{A}\Delta t})^0\mathbf{B} \,\middle|\, (e^{\mathbf{A}\Delta t})^1\mathbf{B} \,\middle|\, (e^{\mathbf{A}\Delta t})^2\mathbf{B} \,\middle|\, \cdots \,\middle|\, (e^{\mathbf{A}\Delta t})^\infty\mathbf{B}\ \right] \tag{23}$$

$$\mathbf{H} = \boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{Q}} \quad ; \quad \mathrm{Rang}(\mathbf{H}) = n \quad . \tag{24}$$

The two matrices $\boldsymbol{\mathcal{P}} \in \mathbb{K}^{\infty\times n}$ and $\boldsymbol{\mathcal{Q}} \in \mathbb{K}^{n\times\infty}$ contain linear independent rows and columns which build the block-dyadic structure. The matrices although possessing an infinite dimension still occupy a finite rank n.

Kalman [3] introduced this process into the systems theory by means of linear algebra. If the rank of the matrices corresponds to the systems order n, then the linear transformations are unique.

The determination of the numerical model parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ is based on the special features of the hankel matrix, the calculation is shown in the following text.

INFLUENCE COEFFICIENTS : If the parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ of a realization $\boldsymbol{\mathcal{S}}\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ for a synthesis or an analysis model are known, then it is still possible to change the state space by means of a similarity transformation. This is the reason why a direct physical interpretation of the state space realization, as presented in the introductory mechanical example, is only possible in exceptional cases.

In engineering Green's function is well known as a solution to the bending differential equation and as such represents the solution in form of an integral equation. The flexibility coefficients δ_{ij} which appear as discrete values in Green's solution are known in civil engineering in the form of influence lines.

The weighting matrix $\mathbf{T}_{(t,\sigma)}$ when seen as an integral solution to the vector differential equation of the first order is comparable to Green's function.

$$\mathbf{y}_{(t)} = \int_{t_0}^{t} \mathbf{T}_{(t,\sigma)}\ \mathbf{u}_{(\sigma)}\, \mathrm{d}\sigma \tag{25}$$

With the aid of the spectral decomposition of the weighting matrix it is possible to gain direct access to a physical interpretation of a realization. After converting the fundamental matrix into a simple Jordan canonical structure

$$e^{\mathbf{J}(t-t_0)} = \mathbf{P}_{\mathbf{J}}^{-1} e^{\mathbf{A}(t-t_0)} \mathbf{P}_{\mathbf{J}} \quad , \tag{26}$$

it is then clearly possible, after ordering the eigenfunctions, to convert the realization with the state space transformation

$$\mathbf{P}_{(t)} = \mathbf{P}_{\mathbf{J}} e^{\mathbf{J}(t-t_0)} \quad , \quad \mathbf{J} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \tag{27}$$

into the state form without the feedback variables $\boldsymbol{\mathcal{S}}\{0,\ \boldsymbol{\mathcal{Q}}_{(t)}, \boldsymbol{\mathcal{P}}_{(t)}\}$. The components of the weighting matrix $\mathbf{T}_{(t,\sigma)} = \boldsymbol{\mathcal{P}}_{(t)} \boldsymbol{\mathcal{Q}}_{(\sigma)}$ then decouple themselves per system eigenfunction.

$$\begin{aligned} \mathbf{y}_{(t)} = & \quad \boldsymbol{\mathcal{P}}_{(t)} \int_{t_0}^{t} \boldsymbol{\mathcal{Q}}_{(\sigma)}\ \mathbf{u}_{(\sigma)}\, \mathrm{d}\sigma \\ = & \quad \mathbf{C} e^{\mathbf{A}(t-t_0)} \int_{t_0}^{t} e^{\mathbf{A}(t_0-\sigma)} \mathbf{B}\ \mathbf{u}_{(\sigma)}\, \mathrm{d}\sigma \\ = & \sum_{i=1}^{n} \int_{t_0}^{t} \mathbf{f}_{\mathcal{P},i} e^{\lambda_i (t-t_0)} >< e^{-\lambda_i^{*}(\sigma-t_0)} \mathbf{f}_{\mathcal{Q},i} \mathbf{u}_{(\sigma)}\, \mathrm{d}\sigma \\ = & \sum_{i=1}^{n} \mathbf{F}_{dyn,i} \int_{t_0}^{t} e^{\lambda_i (t-\sigma)}\ \mathbf{u}_{(\sigma)}\, \mathrm{d}\sigma \end{aligned} \tag{28}$$

When reordered and multiplied dyadically, the integral equation then is made up of a sum of n constant matrices $\mathbf{F}_{dyn}$ and the corresponding system eigenfunction. The matrices $\mathbf{F}_{dyn,i} \in \mathbb{C}^{p\times q}$, which can be built up dyadically from the vectors $\mathbf{f}_{\mathcal{P},i}$ and $\mathbf{f}_{\mathcal{Q},i}$ for every system eigenfunction e^{λ_i}, form the dynamic coefficients of influence.
The constant dynamic coefficient of influence matrix $\mathbf{F}_{dyn}$ is comparable to the influence lines or δ-coefficients in engineering. The coefficients of influence can be interpreted physically according to the principles of cause and effect. The column j of the $\mathbf{F}_{dyn,i}$ matrix shows the influence asserted at the p output points and how it affects a selected input channel j with an excitation in its system eigenfunction.
In contrast to the parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}$ of a realization, the dynamic coefficients of influence $\mathbf{F}_{dyn,i}$ with the corresponding system eigenfunction e^{λ_i} are physically interpreted characteristic quantities for a mathematical model.

REALIZATION OF DYNAMIC SYSTEMS : The term "realization" is used to describe the determination of the parameters of a dynamic system $\mathcal{S}\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$. The determination of the parameters of an analysis model, the black box, can be deduced by means of a hankel transformation using discrete measurements as is described in the following. If the information on a real measured dynamic system is available in the form of two hankel matrices $\mathbf{H}_1; (\mathbf{Y}_i, i = 0, 1, \cdots)$ and $\mathbf{H}_2; (\mathbf{Y}_i, i = 1, 2, \cdots)$, then it is possible to determine the parameters of an analysis model.
Both hankel matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ can always be constructed from a block dyadic form. In this case, the three matrices $\boldsymbol{\mathcal{P}}_1 - \boldsymbol{\mathcal{P}}_2 - \boldsymbol{\mathcal{Q}}$ with their linear independent rows and columns together form in their entirety a generalized block dyadic structure, which in turn enables the determination of the unknown system parameters $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times q}$ and $\mathbf{C} \in \mathbb{R}^{p\times n}$.

$$\begin{aligned} \mathbf{H}_1 &= \boldsymbol{\mathcal{P}}_1\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{P}}(e^{\mathbf{A}\Delta t})^0\boldsymbol{\mathcal{Q}} \\ \mathbf{H}_2 &= \boldsymbol{\mathcal{P}}_2\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{P}}(e^{\mathbf{A}\Delta t})^1\boldsymbol{\mathcal{Q}} \end{aligned} \tag{29}$$

Because of the global time shift invariability of the system the parameter matrix $e^{\mathbf{A}\Delta t} \in \mathbb{R}^{n\times n}$ is contained in the relationship of the two matrices $\boldsymbol{\mathcal{P}}_1$ and $\boldsymbol{\mathcal{P}}_2$ with their linear independent columns. Furthermore, the parameters $\mathbf{C} \in \mathbb{R}^{p\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times q}$ are to be found in the first p-rows and q-columns of the block dyad $\boldsymbol{\mathcal{P}} - \boldsymbol{\mathcal{Q}}$. It is possible to determine the parameters for a dynamic system $\mathcal{S}\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ by means of the generalized block dyadic structure $\boldsymbol{\mathcal{P}}_1 - \boldsymbol{\mathcal{P}}_2 - \boldsymbol{\mathcal{Q}}$ using two hankel matrices $\mathbf{H}_1$ and $\mathbf{H}_2$.
The generalized singular value decomposition of the two measured hankel matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ makes it possible to determine this sought after dyadic block form. The sought after linear independent rows and columns can be determined by decomposing or factorising the hankel matrix with the normal pairs of singular values and the accompanying dyads.

$$\begin{aligned} \mathbf{H}_1 = \breve{\mathbf{U}}\mathbf{X}^{-1} &= \mathbf{U}\boldsymbol{\Sigma}_{\mathbf{H}_1}\mathbf{X}^{-1} = \boldsymbol{\mathcal{P}}_1\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{P}}(e^{\mathbf{A}\Delta t})^0\boldsymbol{\mathcal{Q}} \\ \mathbf{H}_2 = \breve{\mathbf{V}}\mathbf{X}^{-1} &= \mathbf{V}\boldsymbol{\Sigma}_{\mathbf{H}_2}\mathbf{X}^{-1} = \boldsymbol{\mathcal{P}}_2\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{P}}(e^{\mathbf{A}\Delta t})^1\boldsymbol{\mathcal{Q}} \end{aligned} \tag{30}$$

The parameters of the realization are then known with the exception of a possible similarity transformation.

$$\begin{aligned} \mathbf{C} &= \boldsymbol{\mathcal{P}}_{(1:p,1:n)} = \breve{\mathbf{U}}_{(1:p,1:n)} \\ \mathbf{B} &= \boldsymbol{\mathcal{Q}}_{(1:n,1:q)} = \mathbf{X}^{-1}_{(1:n,1:q)} \\ e^{\mathbf{A}\Delta t} &= \boldsymbol{\mathcal{P}}_1^{\dagger}\boldsymbol{\mathcal{P}}_2 = \breve{\mathbf{U}}^{\dagger}\breve{\mathbf{V}} \end{aligned} \tag{31}$$

SEPARATION OF SIGNAL AND NOISE : The input-output values for the analysis model must be available in digital form. Here it is important to distinguish between two ratios of accuracy. Simulated and real measurement must be classified by means of a different numerical quantization.
Simulated values are available in the digital computer precision of $\gamma = 10^{-15}$ and are directly accessible for the calculation of a black box realization. In practice sampled data are full of errors, because of the inaccuracy in measurement [1] [4]. The so called inaccuracy of the measurement can be described in general as white noise. The measured date are each correct to within an approximately A/D-converter precision of $\gamma = 10^{-3}$.
The realization algorithm when based on the generalized singular value decomposition comprises a joint determination of the rank of the two matrices to be factorisied, a fact which is dependent on the precision of the digital computer. When analyzing with measurements of a high inaccuracy it is therefore necessary to insert an external rank determination beforehand, which can also be interpreted as a separation of signal and noise.

$$\hat{\mathbf{H}}_{measurement} = \mathbf{H}_{signal} + \hat{\mathbf{H}}_{noise} = \begin{bmatrix} \hat{\mathbf{H}}_1 \\ \hat{\mathbf{H}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1' \\ \mathbf{V}_2' \end{bmatrix} \quad (32)$$

The principle singular value curve shows the significant differences between the singular values, which are assigned to the signals and those of the noise. The limit is to be found approximately by $\sigma_1\gamma$, which is the maximal singular value multiplied by the numerical quantization.

The curve of singular values, selection of principal values.

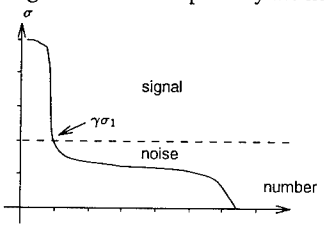


Figure 3: Singular values of the hankel matrix

The complete hankel matrix is reconstructed after the selection of the signal has been made by means of the significant singular values which represent the amount of energy present in the signal to be measured.

$$\mathbf{H}_{signal} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1' \\ 0 \end{bmatrix} \quad (33)$$

The calculation is reduced if the redundant information in the two hankel matrices is taken into consideration.

DAMAGE DETECTION : At this point the methodology of using the dynamic coefficients of influence in order to localize a change in the system [5] together with the state space realization as a method of analysis are briefly outlined. An experimental model is used

to demonstrate the suitability of the methodology in real measured systems. The laboratory model is approximately 2.5 meters long and is made up of two bars with square cross section of steel joined together by means of soft steel bars. The real model was measured and four accelerator sensors were attached at intervals to a length of beam, shown in figure 4.

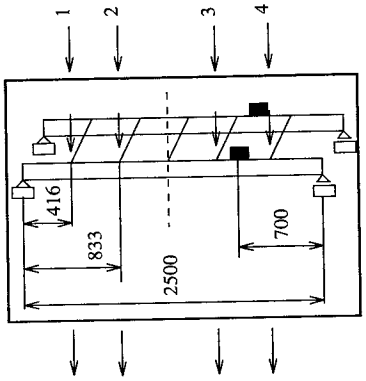


Figure 4: Principle sketch of measuring and location of weight

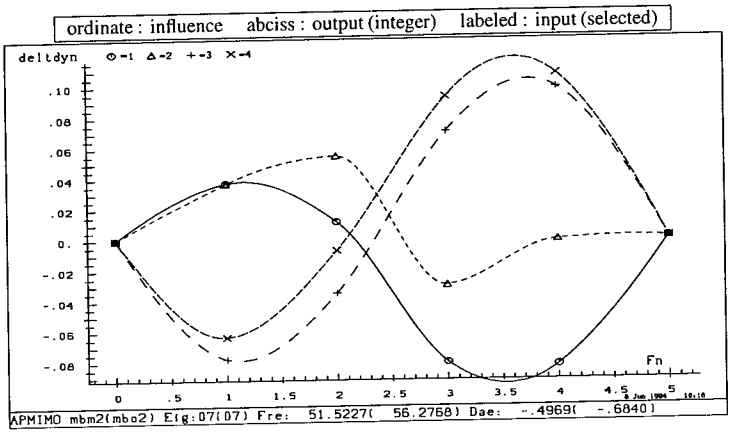


Figure 5: Difference of influence coefficients

An extra weight of 10 % of the overall weight was used to change the dynamic system. The pictures show examples of the difference in the most significant coefficients of influence on the corresponding system eigenfunctions. The lines of influence of the basic eigenfrequency do not allow us to draw conclusions as to the localization of change in the system. This peculiarity is due to the low number of measurement channels.

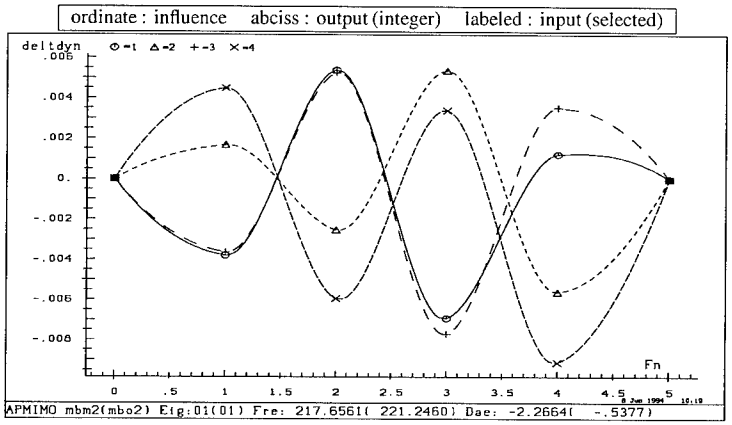


Figure 6: Difference of influence coefficients

The higher eigenfrequencies (figure 5 and figure 6) and their lines of influence point to a change in the structure between the input channels three and four. Taking into consideration the consistent forms of the lines of influence (figure not shown) at the third of the eigenfrequency, it can be deduced that the change in the structure must be located in a form node. Overall it can be said that the lines of influence enable a clear localization of the actual point of change in the system, namely between input channel three and four.

BIBLIOGRAPHY :

[1] G. Golub and C. Loan van. *Matrix Computations.* Johns Hopkins University Press, Baltimore London, 1989.

[2] T. Kailath. *Linear Systems.* Prentice Hall, 1980.

[3] R. E. Kalman. *Mathematical Description of Linear Dynamical Systems.* SIAM J. Control Vol. 1 No. 2, 1963.

[4] A. Lenzen. *Untersuchung von dynamischen Systemen mit der Singulärwertzerlegung – Erfassung von Strukturveränderungen.* Ruhr Universität Bochum, Institut für Mechanik, 1994.

[5] A. Lenzen and H. Waller. *Damage Detection by System Identification.* Archive of Applied Mechanics, Springer, Vo. 66, 1996.

[6] C. Loan van. *Generalizing the Singular Value Decomposition.* SIAM J. Num. Anal. Vol. 13 No. 1, March, 1976.

[7] C. C. Paige. *Computing the Generalized Singular Value Decomposition.* SIAM J. Sci. Stat. Comp. Vol.7 No. 8, October, 1986.

[8] L. A. Zadeh and C. A. Desoer. *Linear System Theory.* Krieger New York, 1979.

DYNAMICS OF MULTIMODAL ROTORS WITH TRANSVERSE CRACKS

Paul Goldman
Agnes Muszynska
Donald E. Bently
Bently Rotor Dynamic Research Corp.
1711 Orbit way, Bldg.1, Minden, Nevada 89423

Abstract: This paper provides insight into the lateral response of a cracked multimodal rotor during the transient processes of startups or shutdowns. The transverse crack on the rotor is treated as a structure singularity which provides an additional local flexibility and mass reduction. The influence of it on the global rotor behavior is estimated. The rotor lateral synchronous, 2x and constant response components are investigated from the mode shape standpoint. The sensitivity of 1x mode shapes to the crack has been proved. The results are formulated as suggestions for vibration diagnostics improvements.

Key Words: Multimodal rotor, crack, diagnostics, lateral vibration.

INTRODUCTION: The continuing interest of scientists and engineers in the dynamic behavior and early detection of the rotors with cracks is stimulated by a relatively high frequency of crack occurrence. According to Bently, Muszynska [1] during a period of 10 years at least 28 rotor failures due to cracks were documented in the USA power industry.

The topic of cracked rotor response has been treated in many papers. Dimentberg [3] was apparently the first to report the effect of the rotating stiffness asymmetry on the shaft lateral vibration. Henry and Okah-Avae [8] performed a computer study where the weight of the rotor was reported to be responsible for the 2x resonance. Mayes and Davis [10] studied the behavior of a cracked shaft model which took into account opening and closing of the crack as a stiffness step-function. Gasch [5] modeled the breathing crack by a spring-loaded hinge, and performed a computer simulation which showed a subharmonic resonance. Grabowski [7] used a modal approach to the problem. He developed a theoretical model of the crack mechanism, which showed a good correlation with static experiments, and used it in a dynamic model of the rotor. Inagaki, Kanaki and Shiraki [9] applied the transfer matrix method. They modeled the breathing crack as a step-function for the bending moment and applied the Fourier series expansion to find the solution. Muszynska [11] considered both gaping and breathing cracks and investigated the interaction between rotating and stationary stiffness asymmetries. Nelson and Nataraj [12] investigated analytically the spectrum change due to the nonlinearities introduced by cracks. Bently and Muszynska [1] emphasized the importance of the observation of the rotor 1x and 2x filtered response vector changes for early crack detection, including not only amplitudes, but also phases. A significant number of papers were published on the finite element modeling of the cracked shafts, for example: Chen and Wang [2], Dirr and Schmallhorst [4].

As it was shown in numerous publications by Bently and Muszynska, in 70% of successful early crack detections cases the most important symptom of the crack is a change of rotor 1x response vector (amplitude and/or phase). The remaining 30% exhibit changes in 2x response. The insufficient knowledge on the impact the crack has on the

rotor mode shapes stimulated the study presented below. This paper is a continuation of the paper by the same authors [6].

MATHEMATICAL MODEL: The goal of this study is to evaluate the multimodal cracked shaft synchronous response behavior during startup or shutdown for further use in the vibrational diagnostics technique. In order to accomplish this task a simple rotor train supported at the ends in two bearings is considered. The rotor dynamics is described in the stationary coordinate system by lateral horizontal $x(s)$ and vertical $y(s)$ displacements of the rotor cross-section center with axial coordinate s, and in coordinates rotating with the rotor by the corresponding lateral displacements $\xi(s)$ and $\eta(s)$. The rotor is weakened by a transverse crack at the axial location s_C in the direction of rotating axis ξ. Since the major dynamic effect of the crack on the rotor vibrational response comes from the rotating stiffness asymmetry, the influence of the breathing crack is neglected. The crack therefore can be described by the different geometric inertia moments about the rotating axes I_ξ and I_η. If the rotary inertia and shear stresses are neglected, the equations for the uniform rotor lateral response can be written as follows:

$$\rho(s)\ddot{x}+E\frac{\partial^2}{\partial s^2}\left(\frac{I_\xi+I_\eta}{2}\frac{\partial^2 x}{\partial s^2}\right)+E\frac{\partial^2}{\partial s^2}\left(\frac{I_\xi-I_\eta}{2}\frac{\partial^2 x}{\partial s^2}\right)\cos 2\Omega t+E\frac{\partial^2}{\partial s^2}\left(\frac{I_\xi-I_\eta}{2}\frac{\partial^2 y}{\partial s^2}\right)\sin 2\Omega t=Q_x$$

$$\rho(s)\ddot{y}+E\frac{\partial^2}{\partial s^2}\left(\frac{I_\xi+I_\eta}{2}\frac{\partial^2 y}{\partial s^2}\right)-E\frac{\partial^2}{\partial s^2}\left(\frac{I_\xi-I_\eta}{2}\frac{\partial^2 y}{\partial s^2}\right)\cos 2\Omega t+E\frac{\partial^2}{\partial s^2}\left(\frac{I_\xi-I_\eta}{2}\frac{\partial^2 x}{\partial s^2}\right)\sin 2\Omega t=Q_y \qquad (1)$$

where Q_x , Q_y are the axial distributions of the radial forces projected on x and y directions, correspondingly. They include distributed unbalance, constant radial sideload, damping forces, and distributed fluid forces. The moments of inertia I_ξ and I_η about the corresponding rotating axes are equal to each other everywhere except at the crack axial location. The appearance of a crack also affects the local linear density ρ of the rotor. The affected parameters relate to those of the undamaged rotor as follows:

$$I_\xi=I_0(s)\left[1-\delta\left(\frac{s-s_c}{L}\right)R_\xi\right],\ I_\eta=I_0(s)\left[1-\delta\left(\frac{s-s_c}{L}\right)R_\eta\right],\ \rho=\rho_0(s)\left[1-\delta\left(\frac{s-s_c}{L}\right)M\right] \qquad (2)$$

where $\delta(\ )$ is a function of Dirac. The expressions (2) can be obtained if the crack is substituted by a beam element with the length Δs and moments of inertia $I_{\xi,\eta}=I_0(s_c)\left(1-\frac{L}{\Delta s}R_{\xi,\eta}\right)$, with linear density $\rho=\rho_0(s_c)\left(1-\frac{L}{\Delta s}M\right)$, and then the corresponding limits, when $\Delta s\to 0$, are considered.

Note that if there is no crack on the rotor then $I_\eta = I_\xi = I_0(s), \rho = \rho_0(s)$ for any axial coordinate s. The system for this case becomes symmetric, and the normal modes $\chi_k(s)$ can be easily introduced. m_k, D_k, ν_k are the modal mass, damping, and natural frequency of the undamaged rotor respectively. The time dependent part $Q_1(s)exp[j(\Omega t+\alpha(s))]$ of the distributed lateral load complex vector $Q_x + jQ_y$ represents distributed unbalance. The constant component is a distributed radial sideload. The component proportional to the velocity is damping force density, and the component perpendicular to the displacement is a distributed fluid force. The relation

$$\int_0^L \chi_k(s)\left[Q_x + jQ_y\right]ds = -D_k(\dot{x}_{0k} + j\dot{y}_{0k}) + m_k e_k \Omega^2 \exp(j\varphi_k) - jP_k + j\Omega D_k \lambda_k (x_{0k} + jy_{0k}) \quad (3)$$

describes the modal damping D_k, unbalance e_k ($\varphi_k = \Omega t + \alpha_k$), radial side load P_k, and Fluid Circumferential Average Velocity Ratio λ_k. For simplicity the modal radial side load force P_k is assumed to be vertical, downward. The 1x rotor lateral response vector $x^{(1)}(s,t) + jy^{(1)}(s,t)$ represents the system reaction to the distributed unbalance force. In rotating coordinates it becomes constant.

$x^{(1)}(s,t) + jy^{(1)}(s,t) = \exp(j\Omega t)\left[\eta^{(1)}(s) + j\xi^{(1)}(s)\right]$ and can be expressed for the undamaged rotor in the following form:

$$\begin{pmatrix} \eta^{(1)} \\ \xi^{(1)} \end{pmatrix} = \sum_{k=1}^{\infty} \begin{pmatrix} \chi_k(s) & 0 \\ 0 & \chi_k(s) \end{pmatrix} \begin{pmatrix} \eta_k^{(1)} \\ \xi_k^{(1)} \end{pmatrix}, \quad \eta_k^{(1)} + j\xi_k^{(1)} = \frac{m_k e_k \Omega^2}{SDS_k} \exp(j\alpha_k), \quad (4)$$

$$SDS_k = m_k\left(\nu_k{}^2 - \Omega^2\right) + j\Omega D_k(1 - \lambda_k)$$

where $\eta_k^{(1)}, \xi_k^{(1)}$ are the components of the k-th mode 1x lateral displacement in rotating coordinates, SDS_k is a synchronous dynamic stiffness of the k-th mode. The modal functions $\chi_k(s)$ of an undamaged rotor can be used to simplify Eqs. (1) for the cracked rotor. In rotating coordinates they can be reduced to the following system:

$$\theta_k'' + 2\theta_k'(\varsigma_{ka}\omega_{ka} + j) + \theta_k\left[\omega_{ka}{}^2 - 1 + 2j\varsigma_{ka}\omega_{ka}(1 - \lambda_k)\right] + \omega_{ka}{}^2 q_{ka}\bar{\theta}_k = e_{ka}e^{j\alpha_k} + \vec{E}_{ka}e^{-j\Omega t} +$$
$$+ \sum_{r=1, r\neq k}^{\infty} \omega_{ra}\omega_{ka}\sqrt{\frac{m_{ra}}{m_{ka}}}a_{kr}\left(\frac{R_\xi + R_\eta}{2}\theta_r - \frac{R_\xi - R_\eta}{2}\bar{\theta}_r\right) \quad (5)$$

where:

$$\Delta_k = \frac{R_\eta + R_\xi}{2}\frac{EI_0(s_c)L}{m_k\nu_k{}^2}\left(\frac{d^2\chi_k}{ds^2}\right)^2\Bigg|_{s=s_C}, \quad \beta_k = \frac{\rho(s_c)L\chi_k(s_c)}{m_k}, \quad q_k = \frac{R_\xi - R_\eta}{2}\frac{EI_0(s_c)L}{m_k\nu_k{}^2}\left(\frac{d^2\chi_k}{ds^2}\right)^2\Bigg|_{s=s_C}$$

$$\omega_{ka}{}^2 = \omega_k{}^2\frac{1-\Delta_k}{1-\beta_k M}, \quad a_{kr} = \frac{EI_0(s_c)L}{\sqrt{m_k m_r(1-\Delta_k)(1-\Delta_r)}\nu_r\nu_k}\left(\frac{d^2\chi_k}{ds^2}\frac{d^2\chi_r}{ds^2}\right)\Bigg|_{s=s_C}, m_{ka} = m_k(1-\beta_k M),$$

$$\varsigma_{ka} = \frac{D_k}{2m_k\nu_k\sqrt{(1-\Delta_k)(1-\beta_k M)}}, \quad q_{ka} = \frac{q_k}{1-\Delta_k}, \quad E_{ka} = \frac{P_k}{m_k\Omega^2(1-\beta_k M)}, \quad e_{ka} = \frac{e_k}{1-\beta_k M}, \quad (6)$$

$$\omega_k = \frac{\nu_k}{\Omega}, \quad \tau = \Omega t, \quad \frac{d}{d\tau} = ', \quad x_k + jy_k = (\eta_k + j\xi_k)\exp(j\tau), \quad \theta_k = \eta_k + j\xi_k, \quad \bar{\theta}_k = \eta_k - j\xi_k$$

Here instead of assumed vertically directed modal radial load $(-jE_{ka})$ a generalized complex vector $\vec{E}_{ka}$ is used. Note that nondimensional coefficients a_{kr} describes the influence of the mode r on the mode k due to the crack, and depend only on the crack axial location. The system of equations (5) includes the crack parameters M, R_η and R_ξ, which are functions of the crack depth and the geometry of the rotor cross-section at the crack location.

Equations (5) are linear and allow for a forced solution $\theta_k = \theta_{k0} + \frac{R_\xi - R_\eta}{2}\theta_{k1}e^{j\Omega t} + \theta_{k2}e^{-j\Omega t}$ where θ_{k0} represents the 1x vector response of the cracked rotor in stationary coordinates, $\frac{R_\xi - R_\eta}{2}\theta_{k1}$ is a 2x vector, and θ_{k2} is a constant component. Balancing the corresponding terms of Eq. (5), the following equations can be derived for different rotor response components:

$$\theta_{ka}\left[\omega_{ka}{}^2 - 1 + 2j\varsigma_{ka}\omega_{ka}(1-\lambda_{ka})\right] + \omega_{ka}{}^2 q_{ka}\bar{\theta}_{k0} = e_{ka}e^{j\alpha_k} + \sum_{r=1,r\neq k}^{\infty}\omega_{ra}\omega_{ka}\sqrt{\frac{m_{ra}}{m_{ka}}}a_{kr}\left(\frac{R_\xi + R_\eta}{2}\theta_{r0} - \frac{R_\xi - R_\eta}{2}\bar{\theta}_{r0}\right) \quad (7)$$

$$\theta_{k1}\left[\omega_{ka}{}^2 - 4 + 2j\varsigma_{ka}\omega_{ka}(2-\lambda_{ka})\right] + \omega_{ka}{}^2 a_{kk}\bar{\theta}_{k2} = \sum_{r=1,r\neq k}^{\infty}\omega_{ra}\omega_{ka}\sqrt{\frac{m_{ra}}{m_{ka}}}a_{kr}\left(\frac{R_\xi + R_\eta}{2}\theta_{r1} - \bar{\theta}_{r2}\right) \quad (8)$$

$$\theta_{k2}\left(\omega_{ka}{}^2 - 2j\varsigma_{ka}\omega_{ka}\lambda_k\right) + \left(\frac{R_\xi - R_\eta}{2}\right)^2\omega_{ka}{}^2 a_{kk}\bar{\theta}_{k1} = \vec{E}_{ka} + \sum_{r=1,r\neq k}^{\infty}\omega_{ra}\omega_{ka}\sqrt{\frac{m_{ra}}{m_{ka}}}a_{kr}\left[\frac{R_\xi + R_\eta}{2}\theta_{r2} - \left(\frac{R_\xi - R_\eta}{2}\right)^2\bar{\theta}_{r1}\right] \quad (9)$$

1X ROTOR RESPONSE COMPONENT: Note that Eqs. (7), which determine the rotor response synchronous component, include only two crack parameters R_η and R_ξ. Equations (7) allow the expression of changes in the total 1x rotor response through the new modal functions $\chi_k^{(i,j)}(s)$ $(i,j = 1,2)$

$$\begin{pmatrix}\eta^{(1)}\\ \xi^{(1)}\end{pmatrix} = \sum_{k=1}^{\infty}\begin{pmatrix}\chi_k^{(1,1)}, \chi_k^{(1,2)}\\ \chi_k^{(2,1)}, \chi_k^{(2,2)}\end{pmatrix}\begin{pmatrix}u_k^{(1)}\\ w_k^{(1)}\end{pmatrix}, \quad \frac{\chi_k^{(1,1)} - \chi_k}{R_\eta} = \frac{\chi_k^{(2,2)} - \chi_k}{R_\xi} = \mathrm{Re}\sum_{s=1,k\neq s}^{\infty}\Omega^2\omega_{ka}\omega_{sa}\frac{\sqrt{m_{sa}m_{ka}}\chi_s a_{sk}}{SDS_{s,1}},$$
$$\frac{\chi_k^{(1,2)}}{R_\xi} = -\frac{\chi_k^{(2,1)}}{R_\eta} = -\mathrm{Im}\sum_{s=1,k\neq s}^{\infty}\Omega^2\omega_{ka}\omega_{sa}\frac{\sqrt{m_{sa}m_{ka}}\chi_s a_{sk}}{SDS_{s,1}} \quad (10)$$

where $u_k^{(1)}, w_k^{(1)}$ are new modal variables:

$$u_k^{(1)} + jw_k^{(1)} = m_{ka}\Omega^2 e_k \exp(j\alpha_k)/SDS_{k,\mathrm{mod}}, \; SDS_{k,\mathrm{mod}} = SDS_{k,1} + SDS_{k,2}$$
$$SDS_{k,1} = m_{ka}\Omega^2\left[\omega_{ka}{}^2 - 1 + 2j\varsigma_{ka}\omega_{ka}\right], \; SDS_{k,2} = m_{ka}\Omega^2\omega_{ka}{}^2 q_{ka}\exp 2j\left[\arg(SDS_{k,1}) - \alpha_k\right] \quad (11)$$

The derivation of relations (10)-(11) has been done under the assumption that the parameters of the crack R_η, R_ξ are of the first order of smallness ($R_\eta, R_\xi = O(\varepsilon), \varepsilon << 1$) and the terms of the second order of smallness were neglected. The modification of the modal dynamic stiffness due to the crack (compare Eq. (11) with Eq. (4)), aside from the simple reduction of the corresponding modal natural frequency and effective damping factor, seen in the first term $SDS_{k,1}$, also includes the rotating modal stiffness asymmetry parameter q_k in the term $SDS_{k,2}$, which depends on the angular orientation α_k of the modal unbalance. The modal 1x response, described by the Eqs. (11) is presented in Fig. 1 in Bode and polar plot format for different angular orientations of the modal unbalance. The maximum 1x amplitude through the resonance corresponds to the unbalance orientation $\alpha_k = 3\pi/4$. The minimum 1x amplitude (lower then that of undamaged rotor) is observed at the unbalance orientation $\alpha_k = \pi/4$ (see Fig. 1, case 2). It is important to note that Fig. 1 describes the behavior of true modal coordinates which differ from the directly measured rotor 1x

response. 1x rotor response represents a sum of products of the modal coordinates and modal functions, which are also affected by the crack (compare Eqs. (4) for the undamaged rotor with Eqs. (10) for the cracked rotor). Note that modal functions in the case of undamaged rotor (see Eqs. (4)) are transformed into 2x2 modal matrices. The modal matrices represent corresponding modal functions.

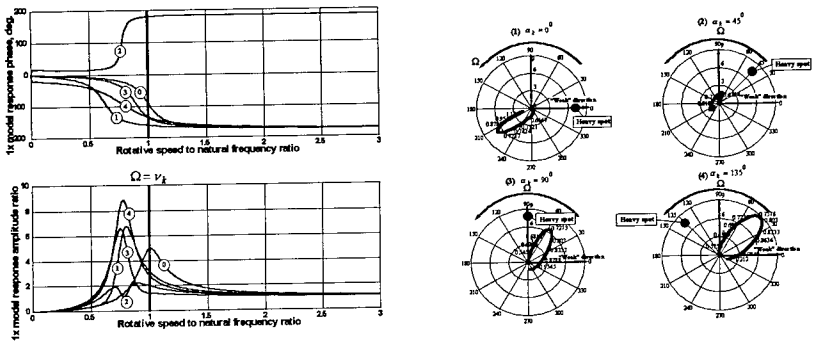


Figure 1. Nondimensional modal 1x response $\left(u_k^{(1)} + jw_k^{(1)}\right)/e_k$ for: undamaged rotor (0) and four different modal unbalance orientations for cracked rotor: 1) $\alpha_k = 0$, 2) $\alpha_k = 45°$, 3) $\alpha_k = 90°$, 4) $\alpha_k = 135°$ $\varsigma_k = 0.2$, $\Delta_k = 0.4, q_k = 0.2$, $M = 0, \lambda_k = 0.48$

2X AND CONSTANT ROTOR RESPONSE COMPONENTS: Equations (8), (9) describe 2x and constant components of the rotor response and are investigated here. A small parameter ε which indicates the magnitude of the major crack parameters ($R_\eta, R_\xi = O(\varepsilon), \varepsilon << 1$) allows the presentation of the solution of Eqs. (8), (9) as a series:

$$\left(\frac{R_\xi - R_\eta}{2}\right)\theta_{k1} = \left(\frac{R_\xi - R_\eta}{2}\right)\left(\theta_{k1}^{(0)} + \varepsilon\theta_{k1}^{(1)} + \varepsilon^2 \ldots\right),\ \theta_{k2} = \theta_{k2}^{(0)} + \varepsilon\theta_{k2}^{(1)} + \varepsilon^2\theta_{k2}^{(2)} + \varepsilon^3 \ldots \tag{12}$$

Substituting Eqs. (12) into Eqs. (8), (9) and balancing the terms of the same order of smallness the following expressions for the terms of the series (12) can be obtained:

$$\theta_{k1}^{(0)} = -\frac{\omega_{ka}}{\omega_{ka}{}^2 - 4 + 2j\varsigma_{ka}\omega_{ka}(2-\lambda_k)}\left[\frac{a_{kk}\bar{E}_{ka}^*}{\omega_{ka} + 2j\varsigma_{ka}\lambda_k} + \sum_{r=1, r\neq k}^{\infty}\sqrt{\frac{m_{ra}}{m_{ka}}}\frac{a_{kr}\bar{E}_{ra}^*}{\omega_{ra} + 2j\varsigma_{ra}\lambda_r}\right],\ \theta_{k2}^{(0)} = \frac{\bar{E}_{ka}}{\omega_{ka}(\omega_{ka} - 2j\varsigma_{ka}\lambda_k)},$$

$$\theta_{k2}^{(1)} = \frac{R_\xi + R_\eta}{2}\frac{1}{\omega_{ka} - 2j\varsigma_{ka}\lambda_k}\sum_{r=1, r\neq k}^{\infty}\omega_{ra}\sqrt{\frac{m_{ra}}{m_{ka}}}a_{kr}\theta_{r2}^{(0)},\ \theta_{k1}^{(1)} = \frac{1}{\omega_{ka}{}^2 - 4 + 2j\varsigma_{ka}\omega_{ka}(2-\lambda_k)}\left[-\omega_{ka}{}^2 a_{kk}\bar{\theta}_{k2}^{(1)} + \right.$$

$$\left. + \sum_{r=1, r\neq k}^{\infty}\omega_{ka}\omega_{ra}\sqrt{\frac{m_{ra}}{m_{ka}}}a_{kr}\left(\frac{R_\xi + R_\eta}{2}\theta_{r1}^{(0)} - \bar{\theta}_{r2}^{(1)}\right)\right],\ \theta_{k2}^{(2)} = \frac{1}{\omega_{ka} - 2j\varsigma_{ka}\lambda_k}\left[-\omega_{ka}a_{kk}\left(\frac{R_\xi - R_\eta}{2}\right)^2\bar{\theta}_{k1}^{(0)} + \right.$$

$$\left. + \sum_{r=1, r\neq k}^{\infty}\omega_{ra}\sqrt{\frac{m_{ra}}{m_{ka}}}a_{kr}\left(\frac{R_\xi + R_\eta}{2}\theta_{r2}^{(1)} - \left(\frac{R_\xi - R_\eta}{2}\right)^2\bar{\theta}_{r1}^{(0)}\right)\right] \tag{13}$$

where $\vec{E}_{ka}^{*}$ is a complex vector conjugated to the vector $\vec{E}_{ka}$. Sufficient accuracy can be achieved if the series (12) in the 2x and constant components of the cracked rotor response are truncated as follows: $\left(\frac{R_\xi - R_\eta}{2}\right)\theta_{k1} \approx \left(\frac{R_\xi - R_\eta}{2}\right)\theta_{k1}^{(0)}$, $\theta_{k2} \approx \theta_{k2}^{(0)} + \varepsilon\theta_{k2}^{(1)} + \varepsilon^2\theta_{k2}^{(2)}$ where $\theta_{k1}^{(0)}$, $\theta_{k2}^{(0)}$, $\theta_{k2}^{(1)}$, $\theta_{k2}^{(2)}$ are determined by the expressions (13). It is important to remember that $\left(\frac{R_\xi - R_\eta}{2}\right)\sum_{k=1}^{\infty}\theta_{k1}\chi_k(s)$ is a 2x filtered rotor response vector at the axial location s, while $\sum_{k=1}^{\infty}\theta_{k2}\chi_k(s)$ is a complex vector of the rotor centerline position at the same axial location.

TWO MODE CRACKED ROTOR RESPONSE: To illustrate the rotor response change due to the crack, an example of a rotor with uniform cross-section ($\rho(s) = const, I_0(s) = const$) and simple rigid supports at the ends is considered. For the purpose of this example all modes higher than the second are neglected. The assumed and calculated major characteristics of the considered rotor are presented in Table 1. The unbalance is provided by the lumped mass with angular orientation α, positioned in the plane with axial coordinate $s_{un} = 0.33L$. The modal radial load is provided by the lumped force $\vec{F} = -jP$, positioned in the same plane. This particular point of the lumped forces application makes the corresponding modal forces for both considered modes equal to each other. The axial location used for the simulated observation point is $s=0.66L$. The rotor lateral response 1x, 2x and constant vectors are calculated, using Eqs. (10), (11), (13) and (14), which are truncated to the first two modes.

Table 1. Parameters of The Cracked Two Mode Rotor

	Natural frequency ν_k	Modal mass m_k	Original modal function $\chi_k(s)$	Modal λ_k	Damping factor ς_k	Coord. of observation	Crack coordinate s_c	Modal coefficients a_{kr}	Crack parameters R_ξ, R_η	Stiffness reduction Δ_k	Stiffness asymmetry q_k
Mode 1	$\pi^2\sqrt{\frac{EI_0}{\rho L^4}}$	ρL	$\sqrt{2}\sin\frac{\pi s}{L}$	0.48	0.1	$0.66L$	$0.2L$	$a_{11} = 0.69$ $a_{12} = 1.12$ $a_{21} = 1.12$ $a_{22} = 1.81$	$R_\xi = 0.3$ $R_\eta = 0.1$	0.14	0.07
Mode 2	$4\pi^2\sqrt{\frac{EI_0}{\rho L^4}}$	ρL	$\sqrt{2}\sin\frac{2\pi s}{L}$	0.48	0.1	$0.66L$	$0.2L$	$a_{11} = 0.69$ $a_{12} = 1.12$ $a_{21} = 1.12$ $a_{22} = 1.81$	$R_\xi = 0.3$ $R_\eta = 0.1$	0.36	0.18

The 1x data in nondimensional form is presented in Fig. 2 in Bode and polar plot format. It shows that the shift of the natural frequency of the second mode is much more

significant than that of the first mode. Depending on angular orientation α, which represent the angle between the unbalance and the shaft "strong" direction, the amplitudes change differently, with maximum at $\alpha=3\pi/4$. Figure 2 represents the data which could be obtained on a real rotor by direct measurements: the 1x response at fixed axial location. It shows the significant differences of 1x rotor response vector of the cracked rotor from that of undamaged rotor mainly in the area of the second natural frequency.

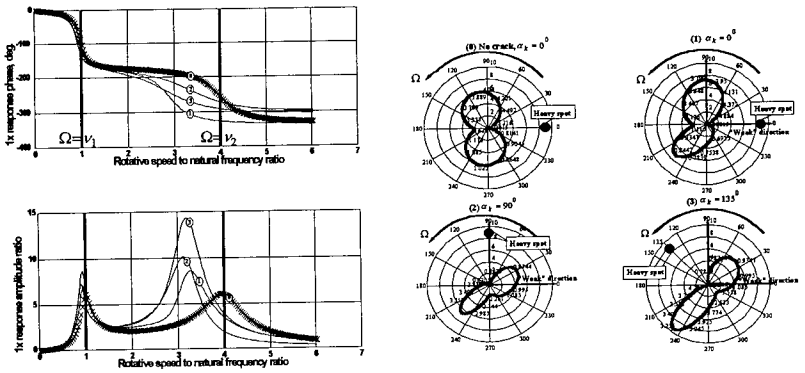


Figure 2. Nondimensional 1x response of the two mode rotor: (0) original response of undamaged rotor, (1)-(3) cracked rotor with the unbalance orientations $\alpha=0$, $\alpha=\pi/2$, $\alpha=3\pi/4$.

That observation is supported by the following modal data. The first mode coordinate (Fig. 3A) and the first mode shape (Fig. 4A) show almost no change due to the crack. At the same time the second mode coordinate (Fig. 3B) and mode shape (Fig. 4B) both are strongly affected. The presented data shows that although the modal coordinate sensitivity to the crack strongly depends on the crack axial position, the first mode shape is almost insensitive to it. On the other hand the second mode is strongly sensitive to the crack, even at speeds far from the second natural frequency. The same conclusion can be reached based on the analysis of the expressions (8), (9). The higher mode shape identification can be used in vibration diagnostics of machinery as a rotor crack detection tool. The existence of a crack at different axial locations reflects on the modification of the 2x resonance frequencies: when the crack axial location is close to the first bearing, the second mode 2x resonance frequency is reduced much more then that of the first mode. The 2x response mode shapes for the two mode cracked rotor are shown in Figure 5 for different rotative speeds. It is obvious that they are close to the original bending modes with certain distortions due to the local flexibility provided by the crack.

The two mode rotor centerline position data during startup or shutdown are presented in Figure 6 at the single axial location and the axial distribution in the form of relative amplitudes and phases. The analysis of Figure 6 for two mode rotor shows two loops in the rotor centerline motion, each in the range of rotative speeds around 1/2 of the corresponding modified natural frequency. Motion along the arc, which follows the second loop is caused by the fluid drag.

CONCLUSIONS:

- The consideration of a crack as a structure singularity leads to the interaction of the originally independent modes. The resulting equations describing the rotor lateral response 1x component can be decoupled by introduction of new modified modes and are easily solved. The results depend on the modal unbalance orientation relative to the crack. The maximum amplitude for the mode occurs when the angle between the modal unbalance and the rotor "strong stiffness" direction is 135°, the minimum amplitude (lower then that of undamaged rotor) occurs when this angle is 45°.

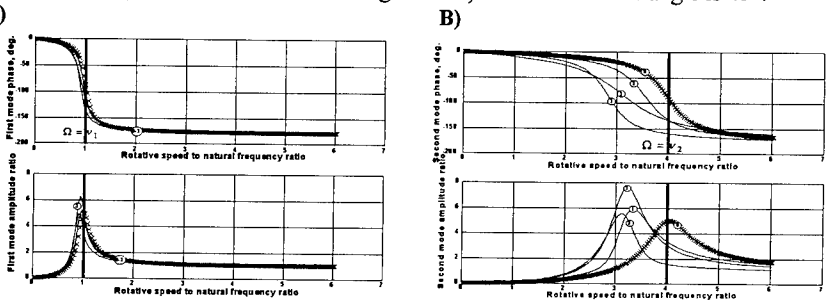


Figure 3. A) first and B) second modes rotor coordinates amplitude and phase: 0) original response of undamaged rotor, (1)-(3) cracked rotor with the unbalance orientations $\alpha = 0$, $\alpha = \pi/2$, $\alpha = 3\pi/4$.

- The example of a two mode cracked rotor shows that the first mode coordinate can be affected by the crack, depending on the axial crack location, but the first mode shape is almost insensitive to the crack. In the case of a crack location close to the support, which is typical for some turbogenerators, the first mode is completely insensitive to the crack.

- The second mode (and in some cases even higher modes) identification is recommended. As a result of crack the mode shapes are significantly changed, and this information can be used as a crack detection tool. From a practical point of view the mode identification can be achieved by the application of the calibration unbalance weight distributed along the rotor based on the original mode of undamaged rotor.

- The 2x cracked rotor response component experiences a resonance each time when the rotative speed coincides with 1/2 of the crack-modified natural frequency of any mode. The phase goes through 180° shift and amplitude peaks at the resonance. The resonance value of the amplitude is roughly proportional to the modal asymmetry factor q_k. The cracked rotor 2x mode shape is approximately the same as corresponding bending modes with slight distortion due to the crack related local flexibility.

- In the case of a multimodal cracked rotor the centerline position exhibits multiple loops around each mode at 1/2 of modified natural frequency with radius roughly

proportional to the modal asymmetry factor in the second power $q_k{}^2$. The sequence of loops is connected with fluid-drag-induced arc. The axial distribution of the rotor centerline position shows that the antinodal point is shifted from the middle of the rotor towards the crack axial position.

A) B)

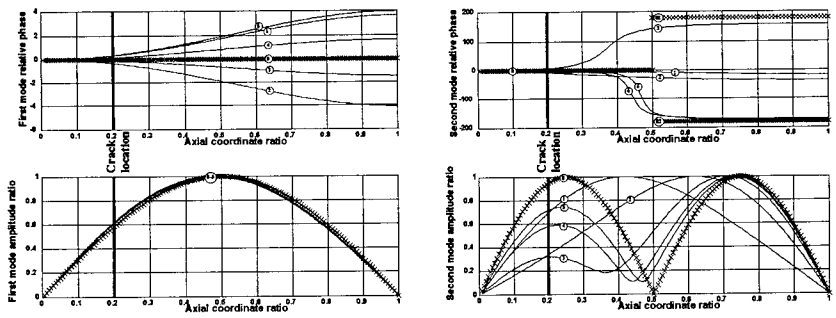


Figure 4. The rotor A) first and B) second modes shapes for: 0) undamaged rotor, in the case of cracked rotor with unbalance orientation $\alpha = 3\pi/4$ for the rotative speeds: 1) $\Omega = 0.7\nu_1$ (below the first natural frequency), 2) $\Omega = 1.1\nu_1$ (above the first natural frequency), 3) $\Omega = 2\nu_1$ (between first and second natural frequencies), 4) $\Omega = 3.6\nu_1$ (below the second natural frequency), 5) $\Omega = 4.1\nu_1$ (above the second natural frequency).

- The appearance of the rotor response 2x predominantly forward component coupled with the described behavior of the rotor centerline position can be used as a crack diagnostic tool for the cases with high rotating stiffness asymmetry. As it was mentioned above it constitutes approximately 30% of all observed cases. The majority of the crack cases has to rely mostly on the 1x component observations.

REFERENCES.

1. Bently, D. E., Muszynska, A., "Detection of Rotor Cracks", Proceedings of 15th Turbomachinery Symposium, Corpus Christi, Texas, Nov. 1986, pp. 129-139.
2. Chen, W. H., Wang, H. L., "Finite element analysis of axisymmetric cracked solid subjected to torsional loadings", Eng. Frac. Mech., vol. 23, No 4, 1986, pp. 705-717.
3. Dimentberg, F. M., "Flexural Vibration of Rotating Shafts", Butterworth, London, 1961.
4. Dirr, B. O., Schmalhorst, B. K., "Crack depth Analysis of a Rotating Shaft by Vibration Measurement", 11th ASME Conf. Vib. Noise, Boston, DE v. 2, 1987.
5. Gasch, R., "Dynamic behavior of a simple rotor with a cross-sectional crack", Vibrations in Rotating Machinery, Inst. of Mech. Eng., London, 1976, pp. 123-128.
6. Goldman, P., Bently, D. E., Muszunska, A.,"The Modal Diagnostics of Rotors with Cracks", ME 96 Congress and Exhibition Proceedings, DE-Vol. 91, Nonlinear Dynamics and Control, ASME 1996.
7. Grabowski, B., "The vibrational behavior of a turbine rotor containing a transverse crack", J. Mech. Des., vol. 102, 1979, pp. 15-19.

8. Henry, T. A., Okah-Avae, B. E., "Vibrations in cracked shafts", Vibrations in Rotating Machinery, Inst. of Mech. Eng., London, 1976, pp. 15-19.
9. Inagaki, Y., Kanaki, H., Shiraki, K., "Transverse Vibrations of a General Cracked Rotor Bearing System", ASME Paper 81-DET-45, 1981.
10. Mayes, I. W., Davies, W. G. R., "The Vibrational Behavior of a Rotating Shaft System Containing a Transverse Crack", I. Mech. E. Conference, Paper C178/76,1976.
11. Muszynska, A., "Shaft Crack Detection", Seventh Machinery Dynamics Seminar, Canada, 1982.
12. Nelson, H. D., Nataraj, C., "The Dynamics of a Rotor System with a Cracked Shaft", ASME J. Vibration, Acoustics, Stress and Reliability in Design, vol. 108, No 2, pp. 189-196.

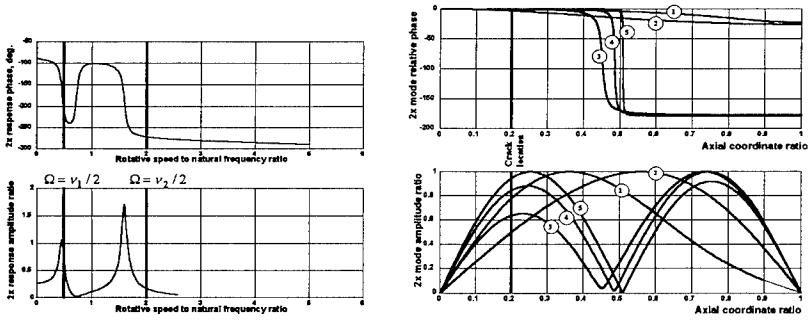


Figure 5. Bode plot of the 2x response vector of the cracked two mode rotor and 2x mode shape in the form of relative phases and amplitudes for different rotative speeds: 1) $\Omega = 0.3\nu_1$ (below 1/2 of the first natural frequency), 2) $\Omega = 0.5\nu_1$ (above 1/2 of the first natural frequency), 3) $\Omega = \nu_1$ (between 1/2 of the first and 1/2 of the second natural frequencies), 4) $\Omega = 1.3\nu_1$ (below 1/2 of the second natural frequency), 5) $\Omega = 2\nu_1$ (above 1/2 of the second natural frequency).

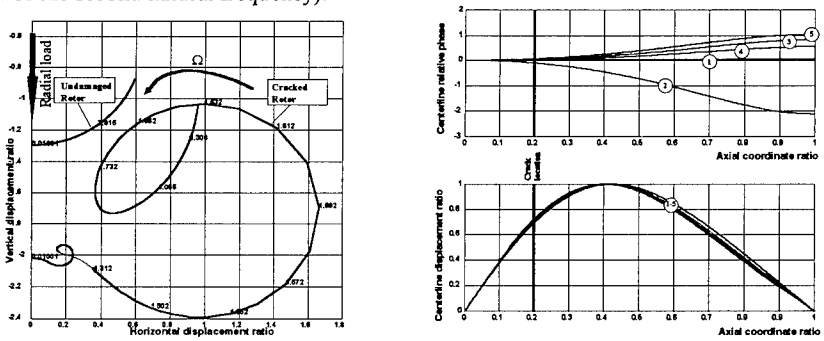


Figure 6. The two mode cracked rotor centerline position and constant displacement axial distribution for different rotative speeds: 1) $\Omega = 0.3\nu_1$, 2) $\Omega = 0.5\nu_1$ 3) $\Omega = \nu_1$, 4) $\Omega = 1.3\nu_1$, 5) $\Omega = 2\nu_1$

CASE HISTORY OF FAILURE DETECTION IN THE DYNAMIC STRUCTURE OF A FACTORY OVERSPEED TEST FACILITY USING "MOVING PICUTURES"

E. M. Scheideler
Siemens AG, KWU Group
Gas Turbine Technology
45473 Mülheim
Germany

Abstract: Fault detection and diagnosis of structure vibration become important. For getting detailed information on the dynamic behaviour of complex dynamic structures however a large amount of measured data is needed. The comparison and reconstruction of the vibration modes of a complet structure from vibration plots as Bode plots and polar plots allown is quite difficult. To solve such problem the use of a simulation program which shows the place and the motion of all measured points in one picture is helpful. Especially the relative motion between the measured points can easily be seen. The "Moving Pictures" method makes it easier to find a potential problem areas.

This paper introduces the procedure of a failure detection for a factory overspeed test facility used for overspeed testing and balancing of gas-turbine rotors. This fault detection was started, because the dynamic behaviour of the facility had changed. As a result, the vibration and the forces at the bearings increased and it was impossible to run a turbine rotor up to their maximum overspeed.

For this vibration diagnosis, 72 vibration pickups at different places and in several directions have been applied. With the method "Moving Pictures" the vibration behaviour of the system (at all measure points) has been detected and visualized at different running speeds. Also the harmonic modes of the system at their natural frequencies have been identified. With the information of the relative motion between the measuring points the problem area and finally the failure was detected. After the repair work was completed, vibration measurements within the problem area were repeated. The evaluation of the measurements showed a very satisfactory vibration behavior of the system, verifying that the failure detection and the repair work were successfull.

Key Words: Overspeed test facility, rotordynamic, vibration measurement, 3D-mesh model

INTRODUCTION: All our gas turbine rotors undergo an overspeed test with a maximum speed of 120 % rated speed before final gasturbine assembly ([1],[2],[3]). This is done in the overspeed test facility. The purpose of this overspeeding is to set all the jointed and shrunk-on components of the rotor so as to minimize any changes in balance during operation at normal running speeds. A fault detection of the factory overspeed test facility was necessary, because the dynamic behaviour of the facility had changed. During running a gas turbine rotor up to its overspeed the vibration level at the bearings became so high, that it was not possible to perform the full overspeed test. The high vibration startet at 107 % rated speed. The method adopted for locating the failure is called "Moving Pictures".

The maine task of the method "Moving Pictures" is the representation of the measured motion in a three dimensional mesh model where the knots of the mesh represent measuring

points. For the better understanding of the measured data of the 3D-mesh model with moving knots can be rotated. Furthermore it is possible to amplify the relative motion independently of the model geometry and to compare the results of different test runs.

If there is a reference measurement the motion of the 3D-mesh from the faulty facility could be compared.

ARRANGEMENT OF THE OVERSPEED TEST FACILITY: The defect factory overspeed facility consists of a outer and inner steel cylinder with poured concrete in between (11m long, outer diameter 6.15m) being supported by an external concrete foundation. Inside this cylinder is the foundation for the cross beams on which the bearing supports are mounted (fig.1). The bearing supports carry the lower half of the bearing via "dynamometer" rods (flat spring) (fig.2, fig.3). The rotor is driven by an outside located electric motor via a gear box and a flexible drive shaft. In the controlroom the 1x vibration components of the rotor support system, the forces, the bearing temperatures and the oil wedge pressure of the bearings are monitored.

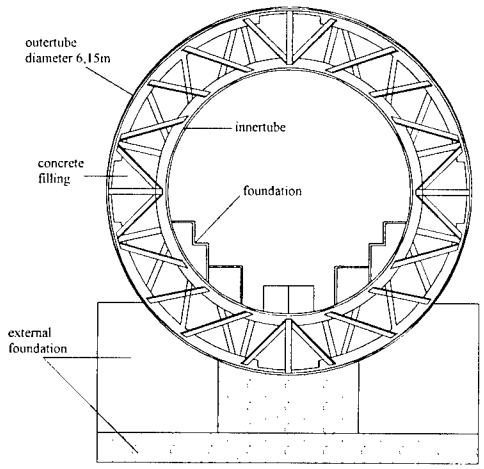


Figure 1: Cross Section of the Factory Overspeed Test Facility

The vibration of the rotor support system is measured with displacement transducers between the bearing lowerhalf and the bottom edge of the cross beam. The measurement is done under 45 degree to the vertical plane. The bearing force is calculated from the vibration which is measured in the same way as described before and with a known stiffness factor. For the calculation of the forces the signal of the displacement measurement is used.

VIBRATION PROBLEMS OF THE OVERSPEED TEST FACILITY: During running a rotor to overspeed, the vibration level and the bearing support forces increased drastically. The induced forces led to excitations of building vibrations within its vicinity. Additionally the bearing temperature increased. The first reaction to solve this problem was to balance the rotor, which did not yield positive results. To be sure, that the rotor itself

is not the cause of the problem, the same rotor was then overspeeded in an other factory overspeed test facility without any problems. With this results further investigations were focused on the different components of the overspeed test facility. The method "Moving Pictures" was selected to conduct these investigations and hence to detect the failure.

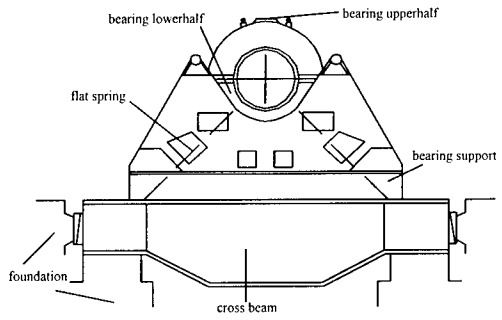


Figure 2: Assembled Rotorbearing Support System

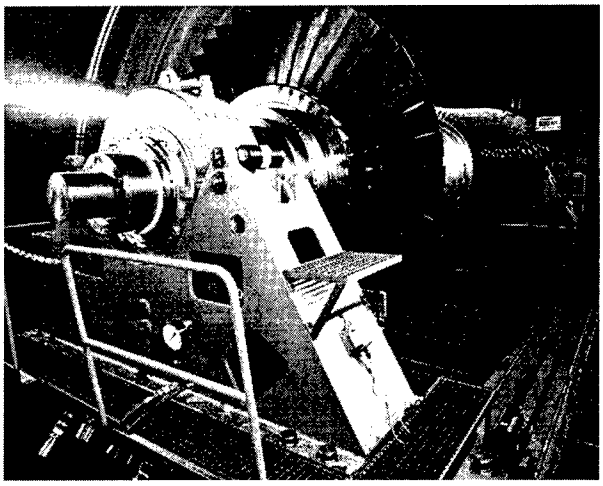


Figure 3: View into the Overspeed Test Facility with a Gas Turbine Rotor

"MOVING PICTURES": The "Moving Pictures" assumes a contemporaneous measurement of several measuring points. Therefore all signals had to be triggered with the same keyphaser probe. The total amount of measurement locations is limited by the hardware. One goal of good "Moving Pictures" is to measure parameters that will provide the best view of the mechanical conditions of the system under investigation.

The main function of the "Moving Pictures" is to view the motion of all measuring points in a three dimensional mesh model. It is possible to show the motion for the time domain or the frequency domain. Also the vibration response to excitation of a modall analyse can be shown.

In the three dimensional model the distance between the knots can be scaled to the real distances and the motion of the knots shown in different scales. There for, the user is able to zoom the motion independently to the scale of the model geometry to get a good feeling of the vibration response of the system.

FAILURE DETECTION OF THE OVERSPEED TEST FACILITY: Several attempts were made with "classical" methods, like looking through the changes around the overspeed facility. Also the manufacturer of the overspeed facility was invited to check the foundation and the bearing support system. During this check the machine foundation and the bearing support system were excited with a shaker . Vibration measurement were taken at different locations. With this approch no failure could be detected.

For the failure detection by "Moving Pictures"a maximum of 72 transducers were used. The electrical signals were changed into digital information with a "diffema 80" system for the time domain and with 8 boxes of the Bently Nevada system ADRE [4] for the frequency domain. Since the number of sensors was limited, it was not possible to measure the motion of the factory overspeed facility at all interesting points at once. So it was decided to split up the measuring program into two parts.

a) First Part of Measurement

With the first test run only the interesting points of the external foundation and the foundation inside the overspeed facility were observed. To be able to verify the compatibility of the measurements from the first and second test run the vibration on the tops of the bearings were detected for both test runs. The reason of this first test run was to make sure, that the foundation and the steel-concrete cylinder have a small vibration response to the excitation of the rotor.

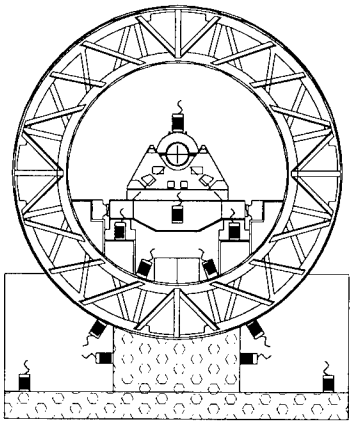


Figure 4: Measuring Points of the Factory Overspeed Test Facility

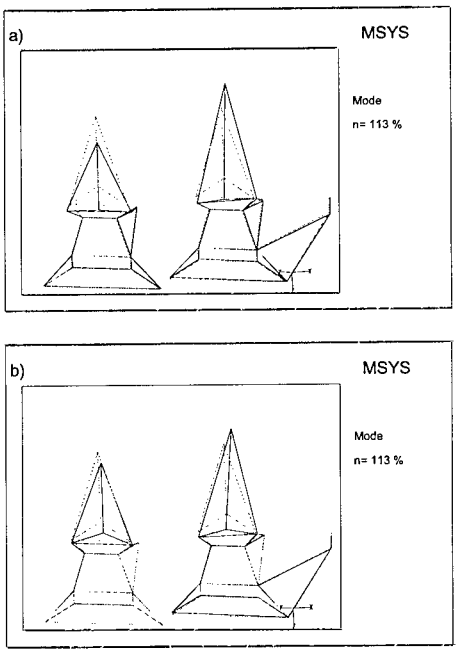


Figure 5: "Moving Pictures" of Overspeed Facility at 113 % Speed

Figure 4 shows the place of the measuring points. The results of the measurement are shown in figure 5a-b. The dashed lines are the 3D-mesh model without any deflection. The solid lines represented the maximum displacement of the components of the overspeed facility. The pictures point out that the foundation, the outer tube and the inner tube do have a vibration behavior in the expected range based on experience. The deflection of the bearing support and the cross beam was however very high. There seems to be a defect in the bearing support system and/ or in the crossbeam. Based on this results the second part of measurements were started.

b) Second Part of Measurement

At the second test run the sensors were installed inside the factory overspeed facility (fig.6). Figure 7a-b shows the deflection of the bearing supports, the crossbeam and the foundation in the tube at 112 % rpm. This figure shows, that the top of the first bearing support oscillates against the crossbeam while the top of the second bearing support oscillates in phase with the crossbeam. The figure shows also, that the crossbeam has a larger deflection as predicted. The polar plots of the proximity probes on the shaft near the bearings show a high vibration amplitude and phase changes at 112% rated speed. It seems that the vibration shifts to a resonance frequency.

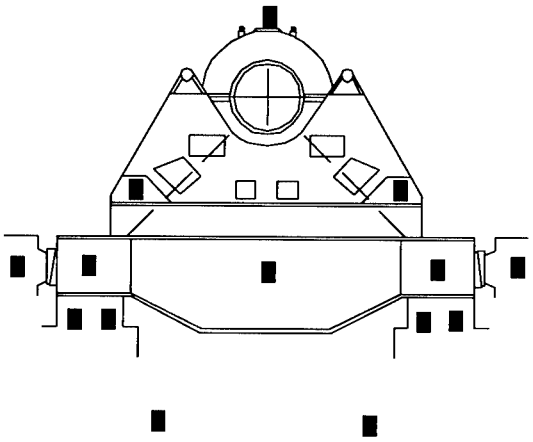


Figure 6: Measuring Points of the Bearing Support System

With the "Moving Pictures" it was possible to view the motion of the factory overspeed facility at any rotating speed. In figure 8 the deflection of the bearing support at 85 % rated speed is shown. The main vibration response moves the bearing support top in axial direction. The amplitude of this motion seems not to be in normally high.

Additionally to the "Moving Pictures", the information of the Bode plot points out, that the vibration signals at higher speed have non linear terms. This could be caused by looseness in the connection between the flat springs and the bearings or between the bearing support and the crossbeam. Because of the above mentioned results the bearing support system were disassembled.

COMPARISON OF THE VIBRATION RESPONSE BEFORE AND AFTER REPAIR OF THE BEARING SUPPORT: When the bearing support was disassembled the following problems were found:

1. The connection between the bearing support and the crossbeam was defect. The screws and exactly their thread were damaged
2. The bolted joint between the flat springs and the edge of the beraing support was damaged
3. The flat springs showed some cracks
4. The axial damping elements were defect

After repairing these failures a reference measurement was performed. The location of the sensors on the bearing support and the crossbeam was not changed.

Figure 9 shows the maximum deflection of the rotor at overspeed (112 %). The bearing tops oscillate against the crossbeam. The vibration of the bearing support is symmetric. The deflection of the crossbeam now was very small.

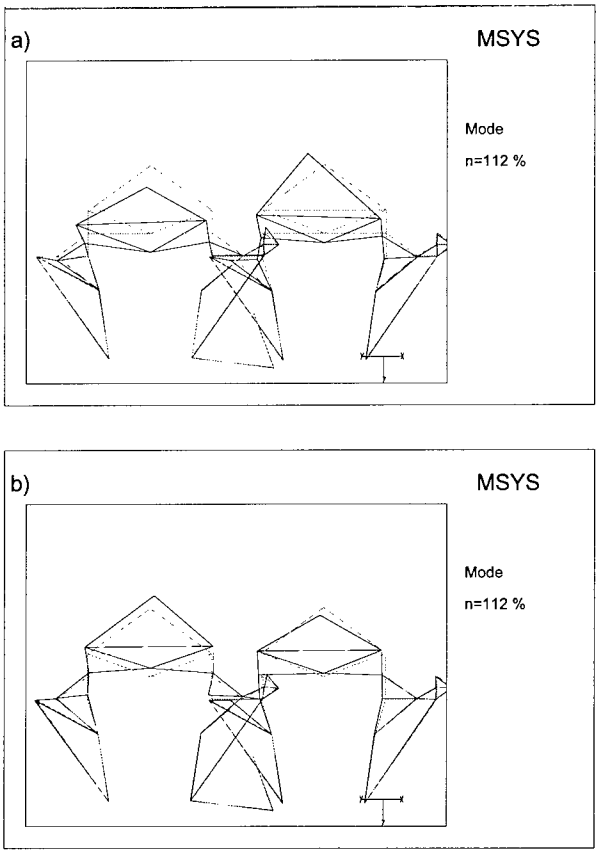


Figure 7: "Moving Pictures" of Bearing Support at 112 % Speed

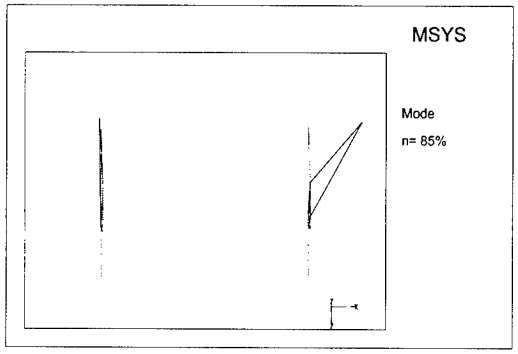


Figure 8: Axial Deflection of Bearing Support at 85 % Speed

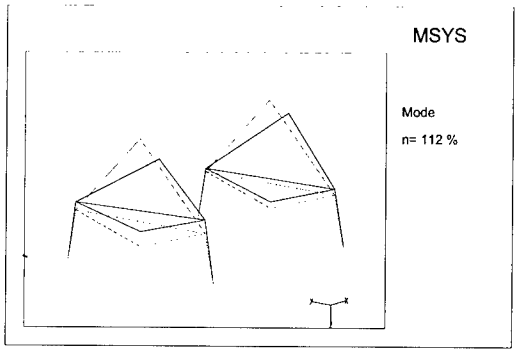


Figure 9: "Moving Pictures" of Bearing Support at 112 % Speed

In the table below the ratio between the vibration amplitudes before and after repair at 113 % rated speed are listed.

bearing top 1		**bearing top 2**	
direction	ratio [%]	direction	ratio [%]
vertical	18,04	vertical	13,84
horizontal	4,69	horizontal	5,41
axial	4,25	axial	7,11

The bearing top vibration level after repair were much smaller than before. From the plot amplitudes versus frequency with the direct signal and the 1x component it is known, that the vibration consists mainly of a linear 1x component, the non linear terms are small.

The vibration values and the vibration behavior meet the standards and are in the range of experience gained during almost 2 decades of testing in this overspeed facility.

CONCLUSION: By using the method - "Moving Pictures" - it was possible to detect the location of the failure in the factory overspeed facility. After repairing the defects it has been shown, that the vibration level to the original conditions decreased. "Moving Pictures" used in cooperation with the classical diagnostic techniques - such as bode plots, waterfall diagrams, polarplots etc.- is a powerfull instrument to localize the defects in complex dynamic structures. If only the classical diagnostic techniques were used it is often difficult to imagine the vibration shape of the structure, and to see which part of the complex structure behave not in the expected range based on the experiences.

REFERENCES:

[1] British Standard 132

[2] DIN 57530 Part 3

[3] API Standard 616

[4] Advanced machinery dynamics cours, Court House, Bently Nevada GmbH

[5] Ehrich, F. F., 1992, Handbook of Rotordynamics, MacGraw hill

[6] Scheideler, E. M., 1996, "Arbeitsbericht KWU/WB/TLS 96/14 , Siemens KWU Germany

[7] Maul-Theet Systeme, 1996, Arbeitsbericht Messungen am Wuchtbunker, Mault-Theet Systeme GmbH, Germany

[8] Janssen, M. J.; Joyce J.S., 1996, 35-Year Old Splined-Disc Rotor Design for Large Gas Turbines, ASME 96–GT–523

ROBUST FAULT DETECTION OF LARGE VIBRATING STRUCTURES BY MEANS OF CONTROL THEORY - SOME PRINCIPAL REMARKS

Dirk Söffker
Safety Control Engineering, University of Wuppertal
Gaußstraße 20, D-42097 Wuppertal, Germany

Abstract: Core of this contribution is the comparison of different fault detection philosophies applied to large vibrating structures. Furthermore ideas are given to overcome application problems of causal fault detection.

Key Words: Observer, Parameter Identification, Signal Analysis, Fault Detection

INTRODUCTION: Reliability and safety aspects are becoming much more important due to higher quality requirements, complicated and/or connected processes and structures. The fault detection approaches to be commonly used in machine- and rotordynamics - in general: of large vibrating structures (LVS) - are based on signal analysis methods. By this way the human knowledge of the behavior of the unfaulty system is used as a base for the comparison with the actual behavior. Applying signal analysis methods (fourier transforms, spectrums etc.) the vibrational behavior will be monitored very well, but has to be interpreted.
The signal based approaches do not use the system knowledge, especially the mechanical parameters of the structure. This available knowledge is typically used by the operating staff interpreting the resulting signal parameters.
Applying methods of modern control theory these problemes can be defused. In this contribution a principal overview about results concerning observers and estimators as methods of the modern control theory will be given. In this context the aim of those approaches is to observe the system behavior, and to detect system changes.
In the last years several methods of modern control theory are specified and applied
- to fault detection ascertaining a failure of sensors/components/faults in/of the system,
- to fault diagnosis determining the existence of specified faults, and
- to fault isolation, which implies the separation of further effects regarded as unknown inputs with respect to the changes caused by the fault in dynamical systems and structures.

Keywords of the last decade are Residual Generator, Decision Maker, Extended Kalman Filter, Parity Equation and Diagnostic Observer /1,3,8,13,14/.
Actual developments are denoted by the consideration / the separation of the influence of modeling errors which includes disturbance decoupling /8/, and also aspects of causal - based fault diagnosis, i.e. the assignment physical fault - monitoring parameter.
Applying signal analysis methods, fault detection leads to the comparison of actual signal values with old values of ordinary behavior, or the comparison to maximum allowed values. These strategies are based on pattern recognition methods. Pattern recognition methods can be applied to compare the effects of the faults found in various experiments

or simulations to those resulting from the considered system.
The mentioned developments are characterized by some indirect assumptions and restrictions which should be noted here: The standard methods of fault detection allow the information condensation of the measurable dynamical behavior to some characteristic values, which are observed. This only implies the use of the outputs of the system. Information about structure and parameter and also of the inputs of the system are not used.

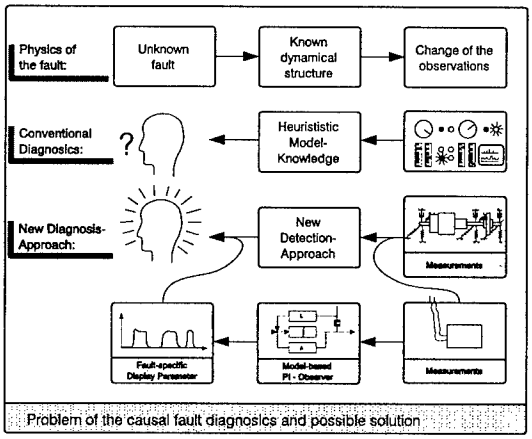


Fig. 1: Procedure of the human fault reasoning process

In fact this contains only a phenomenological study of the system using directly available measurements. Different faults are assumend by different signatures of the characteristic values, but this does not include a causal analysis of the fault or the change itself. The relation between the parameters determined by signal analysis and the physical fault is established by the knowledge and experience of the machine operator, but it should be noted, that this relation often is not unique, but solved by the operators in various practical situations.
The typical reasoning process is given in fig. 1. The experiences are collected as results by doing simulations and experiments, cf. first row. Here starting with given boundary and initial conditions, determined system changes lead to typical results, like signatures. The fault reasoning process turns this direction back. From the observed phenomena it is concluded to the supposed causal reason.
On the other hand several properties of the mentioned strategy make the 'human-decision-making process' itself difficult:

- The relation between the phenomena given by the measurements and the fault is ambiguous, so the results of signal analysis have to be interpreted. Especially in the case of methods, which are not close (enough) to the process itself, the area of possible interpreting errors is wide and depends on the individual human.
- Economical and psychological aspects determine the situation in which the operators have to decide. Proving the existence of a shaft crack by stopping the turbine of a power plant e.g., is a costly procedure. If the non-existence of a crack is definitly obtained, this increases the sensibility for the next decision situation, where the problem appears again.

- Especially in the case of rare faults with high dangerous potential (in atomic power plants or in chemical industry) this 'human decision making process' is very difficult and critical. As a result of these experiences, the challenge is to apply new methods, which lead to a better inner view of the system to be observed. This can be done by applying methods, which use more information about the system itself.

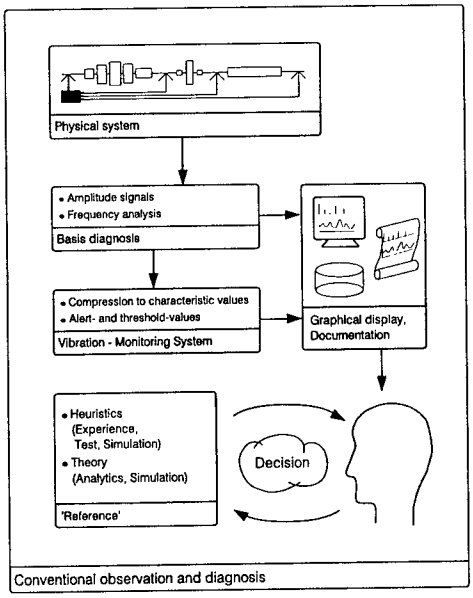


Fig. 2: Conventional observation and diagnosis

The next sections introduce the quantitative model-based approaches briefly and from a principal point of view. Details for further information can be found in the numerous literature.

ANALYTICAL REDUNDANCY METHODS - OBSERVERS: Observers are the base for an analytical redundancy approach for fault detection. Here analytical relationships describing dynamical relations (in contrast to hardware redundancy) are used for detection of changes. The assumed analytical redundant description are represented by the plant (and possibly the failure) model. The interesting criteria are the capability of the scheme concerning isolability, sensitivity and robustness.

Starting with the linear, time - invariant state space description

$$\dot{x} = Ax + Bu + E_d d_d + K_d f_d \quad y = Cx + E_m d_m + K_m f_m \tag{1}$$

with the n-dimensional state vector x, the r-dimensional input vector u, the m-dimensional output vector y, the system matrix A, the input vector B, the measurement matrix C, the distribution matrices E_d, E_m of the unknown inputs d_d and d_m acting on the dynamics d_d and the measurements d_m, the distribution matrices K_d, K_m of the fault influences f_d, f_m which are also of appropiate dimensions, the task to build up an obser-

ver using the given system description with $(A,\ C)$ - observerbility (or the observerbility of an interesting subspace) to estimate $\hat{x}$, or an interesting supspace, respectively.
Differences between observed / estimated and measured behavior expressed by the estimation error

$$e = \hat{x} - x \tag{2}$$

are used as residuals for basic FDI-concepts. The residual analysis by different strategies can be used checking occurence of faults and their location. In this concept it should be noted that the FDI-concepts devide between

- instrument fault detection ($\rightarrow$ sensor fault detection),
- component fault detection ($\rightarrow$ parts of the interconnected system) and
- actuator fault detection ($\rightarrow$ actuators as parts of systems to be controled).

In the context of fault detection schemes for application to LVS vibrating structures only the fault detection idea itself is of interest for the main task of structural and parametric change detection.
The parity space approach uses the observability relation between the measurements (done by sensors) and the mechanical system. So sensor failure leads directly to changes in the observation relations. Using some structural and logic informations a so called parity space vector can be used in finding the failure by considering the vector elements. The success of such approaches based upon a good agreement between the real (unfaulty / fault free) system and the representing model. In the non - corresponding case errors in the error equation occur and problems concerning the differentation between model error and appearing fault appear.
The development of robust residual generator, in which the generated estimations of inner states and measurements are independent of uncertainties are the goals of actual developments. Here, two research directions will be declared briefly:
Unknown Input Observer: The idea of decoupling uncertainties (caused by modeling errors) and change effects through faults from a mathematical exact approach is core of the Unknown Input Observer (UIO) or in general of the robust residual generation. Here a lot of scientific work has been done. Briefly a short introduction is given based on the papers of Frank /4,5/, where also the mathematically exact statements can be found:
Following the description in /4/ the dynamical system

$$\dot{z} \;=\; Fz + Ju + Gy \qquad r = L_1 z + L_2 y \tag{3}$$

with z the t-dimensional subspace of x and the v-dimensional subspace r of the outputs, called residual, is considered. This system is an unknown input fault detection observer of the system (1,2).
The corresponding equation for the estimation error is

$$\dot{e} \;=\; \dot{z} - T\dot{x} \tag{4}$$

$$\dot{e} \;=\; Fz + ju + GCx + GE_m d_m + GK_m f_m - TAx - TBu - TE_d d_d - TK_d f_d \tag{5}$$

with the output relation

$$t = L_1 z + L_2 C x + L_2 E_m d_m + L_2 K_m f_m\,. \tag{6}$$

To fulfill the mentioned robustness requirements exactly, some equations have to be solved:

$$TA - FT \;=\; GC \qquad J = TB \qquad TE_d = 0 \tag{7}$$

$$GE_m = 0 \qquad L_2E_m = 0 \qquad L_1T + L_2C = 0 \ . \tag{8}$$

The error dynamics of the residual results as

$$\dot{e} = Fe + GK_m f_m - TK_d f_d \qquad r = L_1 e + L_2 K_m f_m \ . \tag{9}$$

The conditions therefore are

$$\text{Rank } [TK_d] = \text{Rank } [K_d] \quad \text{and} \quad \text{Rank} \left[\begin{pmatrix} G \\ L_2 \end{pmatrix} K_m \right] = \text{Rank } [K_m] \tag{10}$$

as formulated in (Frank, 1993). In many practical cases these conditions can not be fulfilled. To solve the design procedure to fulfill eqn. (9-14), methods are given in /23/ using the Kronecker canonical representation, in /12/ using the eigenstructure assignment approach. New results are given in /9,10/. The main problem for application to LVS is concerned with the strong assumption that distinguishing p faults and q unknown inputs p+q independent measurements are needed /5/. Because of the large number of modeled elastic degrees of freedom this requirements often can not be fulfilled.

Proportional Intergral Observer: The second approach to be introduced is the PI-observer, as an approximate procedure for decoupling uncertainties and faults. Here the decoupling results not from the observer design itself (as with the UIO), but from the high observer gains. The idea is presented in /21/ and applicated to FDI of mechanical schemes in general /22/.
Assuming that the system with unknown inputs (faults, modeling errors, further external disturbances) is described by

$$\dot{x} = A_o x + Bu + \Delta A(x,u,\mathrm{t})x = A_o x + Bu + \tilde{N}\tilde{f} \ , \tag{11}$$

where A_o describes the unfaulty, nominal system, and $\Delta A(x,u,\mathrm{t})$ (Eq. (19)) describes the influences of system changes due to faults and additional inputs due to external disturbances etc. . Here $\tilde{N}$ in Eq. (20) includes all the external effects acting on the nominal system and is of full rank. If the system (20) is observable and some conditions concerning the number of independent measurements r_1 = rank C, then there exists a PI-Observer introduced in /21/, which estimates the unknown inputs in r_2 input channels. The basic idea of the Proportional Integral Observer (PIO) is to extend the usual Luenberger observer with some degrees of freedom. These degress of freedom are used for the estimations of the unknown inputs and are driven by the integral of the estimation error of the measurements.
Then the idea is that the additional inputs can be divided by

$$\Delta A(x,u,\mathrm{t})x = \tilde{N}\tilde{n} = Nf + N_c f_c \tag{12}$$

where N selects the interesting input channels for the unknown inputs f to be estimated, considering dynamical effects of faults, located to desired inputs selected by N similar to the matrices E_d, K_d of the UIO.
The product $N_c f_c$ describes the remaining inputs of the other input channels. With

$$rank\, N = \mathrm{r}_2 \leq rank\, C = \mathrm{r}_1 \ , \tag{13}$$

applying high gains, this technique is easy to understand.
With $\begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = \rho_1 \begin{bmatrix} \tilde{L}_1 \\ \tilde{L}_2 \end{bmatrix}$ an extended error equation can be written as

$$\frac{1}{\rho_1}\begin{bmatrix} \dot{e} \\ \dot{\hat{f}} \end{bmatrix} = \frac{1}{\rho_1}\begin{bmatrix} A & N \\ 0 & 0 \end{bmatrix}\begin{bmatrix} e \\ \hat{f} \end{bmatrix} - \begin{bmatrix} \tilde{L}_1 \\ \tilde{L}_2 \end{bmatrix} Ce - \frac{1}{\rho_1}\begin{bmatrix} N \\ 0 \end{bmatrix} f - \frac{1}{\rho_1}\begin{bmatrix} N_c \\ 0 \end{bmatrix} f_c \quad .$$

From eq. (23) it follows that

$$Ce = 0 \tag{14}$$

for $\rho_1 \rightarrow \infty$. Differentiating eq. (23) and using (24) gives

$$C\dot{e} = C(A - L_1C)e + CN(\hat{f} - f) + CN_c(\hat{f} - f_c) \quad . \tag{15}$$

Assuming

$$CN = 0 \tag{16}$$

and eq. (24) it can be seen that

$$CAe = 0 \quad . \tag{17}$$

In the same way assuming the observerbility of the complete system it is obtained

$$CA^{\mathrm{i}}e = 0 \quad \mathrm{i} = 0, 1, ..., \mathrm{k} - 1 \quad . \tag{18}$$

Then from (24), (26) and (27) it follows that

$$e = 0 \tag{19}$$

due a rank condition related to the assumption of the observability. Substituting (29) into (23) gives with $\rho_1 \rightarrow \infty$

$$0 = N\hat{f} - Nf - N_c f_c \quad . \tag{20}$$

With $N = \rho_2 \tilde{N}$, $\rho_2 \rightarrow \infty$ eq. (30) gives

$$\hat{f} - f = 0 \quad , \tag{21}$$

i.e. the estimates $\hat{x}$ and $\hat{f}$ of the PIO (12) converge to the system states and the unknown inputs f desired by the matrix N. This result is obtained in presence of other unknown inputs in all remaining states when ρ_1 and ρ_2 tends to infinity. The complete proof is given in /22/. This short introduction shows that the application of the PIO is similar to the application of an UIO for fault detection and isolation in the presence of modeling errors etc. . The conditions for applications are weaker then those of the UIO. Related to the LVS the conditions can be easily fulfilled. The main restrictions are that only displacements (instead of velocities) can be used (eq. 26) and the assumptions given by eq. (22). Related to the practical demand of causal fault detection both approaches have the problem for application of multiplicative faults.

PARAMETER ESTIMATION APPROACHES: Parameter estimation approaches also use the input-output relation modeled by difference- oder differential equations using a set Θ of parameters a_i, b_i for the input - output relation

$$y(k) + ... + a_n y(k - n) = b_0 u(k) + b_1 u(k - 1) + ... + b_m u(k - m) \tag{22}$$

as a scalar difference equation (or a set of equations) between the output y and the input u, with k as the actual discrete time step.
The parameters a_i, b_i do not typically represent the physical parameters p_i of the system, but if the relation

$$\Theta = f(p_i) \tag{23}$$

is unique, an inverse transformation f^{-1} gives the physical parameters. In /15/ a fast backtransformation approach is introduced.
Changes of the physical system are assumed as changes Δp_i, which lead also to changes of the process coefficients a_i, b_i. The idea of FDI using parameter estimation and identification approaches is to observe the parameters (process or physical parameters) using threshold values, some heuristic methods inspired from practice or stochastic decision theory. For solving the estimation process a lot of estimation techniques are available like least squares (LS) or recursive least squares (RLS) techniques. Important is that the assumptions concerning the available measurements φ are

$$det\ E\{\varphi(\mathrm{t})\varphi(\mathrm{t})^T\} \neq 0\,. \tag{24}$$

This includes a high information level concerning the measurements of the system. This can be easily be fulfilled in many practical cases beside elastic structures. Here due to the high number of modeled elastic degrees of freedom this assumption is hard to fulfill. For every degree of freedom (dof) an independent measurement is needed.
Furthermore the robustness properties are similar to observer based methods. Here the structure of the input - output relation (concerning order, linearity, characterics of nonlinearity) also has to be known. An implicit assumption is, that the structure of the system (represented by the structure of difference- /differential equation(s)) does not change due to the fault. One way to overcome these difficulties is to consider only the modal characteristics (Basseville et al., 1993b). It should be noted, that this idea is based on a linear approach (modal characteristics, eigenfrequencies, eigenvectors).

COMBINED APPROACHES: Using the (nonlinear / linear) Extended Kalman Filter /11,17/ for the dynamical model extension of the unfaulty system, a fault model is used, which has to be parametrized. If the nominal system is not exactly known, the same problems as mentioned with analytical redundancy methods appear: faults and modeling errors has to be decoupled. The idea of modeling the fault implies furthermore the more difficult problem to assume the structure and the order of the mechanical description of the fault (same assumptions as for parameter estimations approaches).
In spite of the mentioned assumptions this approach has also been successfully applied to many practical mechanical problems. As example the work of Seibold /19,20/ is mentioned. Here some approaches for crack detection of a rotor, using only the measurements of the bearings, are compared.
The idea of combining the Extended Kalman Filter approach with an unstructured extension (as an integral feedback like the PI-observer approach) is firstly given in /18/ and is called 'Modified Extended Kalman Filter' (MEKF).
In the case that the nominal system (which is the base building up the Kalman Filter) also is not known, a combination of Kalman filtering and parameter identification techniques can be used /6,7/. In this way an adaptive fault detection approach is established, which

avoids problems of modeling errors in a certain way. On the other hand systems changes due to faults can also be adapted as modeling errors.

CRITICAL REMARKS: The most FDI-schemes and approaches have significant advantages concerning special related problems /14/. Several conferences like SAFEPROCESS focus the community to this application field of observers / filters and identifiers. In this contribution the interest is focussed to the application of fault detection schemes to (mechanical) LVS. Fault detection of vibrating structures implies that only the measurements of the vibrating structures can be used beside the system / model information itself.

Criteria / Technique	①	②	③	④	⑤	⑥	⑦
UIO	+	-	+	+	- -	o	+
PIO	+	o	+	+	+	o	+
Parity	+	+	+	+	+	o	+
EKF	+	+ +	+ +	+	+	- -	+
PIA	+	o	-	o	- -	o	+ +
Combined PIA + EKF	-	+	o	o	+	o	o
Signal based approaches	+ +	-	-	- -	+ +	+ +	+ +

Comparison of FDI-Approaches concerning large vibrating structures

Fig. 3: Comparison table of FDI-approaches used for LVS
(Notation: ++ : very good, + : good, o : average, – : very bad, - : bad)

Considerung the advanced fault detection problems of large vibrating structures (examples: cracks in rotating machinery, wings of large aircrafts; changes of space truss structures) the following criteria can be build up:
1. Suitability for On-line-use, **2.** Uniqueness of the relation: System changes - Monitored parameter, **3.** Vicinity of the observed parameter to the interesting, physical problem, **4.** Effort to conclude from the monitored parameter to the physical change, **5.** Expenditure for the necessary measurements, **6.** Exactness of necessary system information and **7.** Practicability concerning the realisation in situ

- Considering criteria 1 (On-line use) the signal-based approaches are the fastest ones, the other approaches are realizable, the combined approach has an additional inner dynamics which needs time.
- Considering criteria 2 (Uniqueness) the Extended Kalman Filter is the best because of the included fault model (costs: criteria 6). Negative results are possible if the output is ambigiuos related to the faults.
- As criteria 3 the vicinity of the outputs to the faults is valuated. If a model is used for

the FDI-approach (like EKF) this leads to good results (costs: criteria 6), if the parameters are unrelated to faults this is negative.

• Considering criteria 4 (Conclusion effort) the EKF get (++) notation, bacause if a model is available for the fault there is nothing to conclude; signal based approaches get (-) notation because the effort depends on the knowledge of the interpreter.

• Considering criteria 5 (Expenditure for the measurements) (++) notation is given if the measurements are easy to realize (signal based approaches) or (−) notation if this is difficult.

• Criteria 6 judges the exactness of the necessary model information (system and fault) for successful application. The EKF get (−) notation because of using typically unknown but necessary fault model, the signal-based approaches get (++) notation because no model is needed (costs: criteria 4).

• Criteria 7 (Practicability of the algorithm itself) judges the effort for the practical engineer to apply and handle the considered approach. If nothing has to be done (++) notation is given, if the algorithm is complicated, needs corrects starting values etc. (o) notation is given.

All of the approaches have advantages and disadvantages. This is obvious comparing EKF and signal-based approaches. The output of EKF is clear, unambigious and unique but strongly model-dependent. The output of signal-based approaches is not unique and has to be interpreted but the application is simple.

The combined parameter identification + EKF approach is not useful because of the adaptation. The PIA can also adapt the fault, if the fault model of EKF is not exact known.

WHAT HAS TO BE DONE IN FDI FOR LVS? The comparison of the different approaches applied to large vibrating structures in fig. 3 shows:

• Approaches easy to apply (criteria 1,5,6,7) (→ Signal-based approaches) have disadvantages concerning to the causality assignment (criteria 2,3,4). To overcome these disadvantages such approaches can be combined with knowledge based techniques like expert systems, but it should be noted that the principal problems depicted in fig. 2 make it difficult to solve the causality requirements for every fault problem (criteria 2,4). This is mainly due to the (for this case poor) information level of the measurements.

• Approaches with an 'easy-to-understand-output' (→ EKF, Parity space approach) (criteria 2,3,4) have disadvantages concerning the assumptions related to the necessary system and fault model information. To overcome these disadvantages the assumption of the knowledge of the fault model has to be canceled. A first approach is done in the work of Seibold et al. (Seibold, Söffker, Fritzen, 1993b). In this case the extension of the EKF is unstructured. Applied to mechanical systems forces and/or torques are estimated. This modified EKF and the PIO are identical, beside the fact that there exist different techniques for the design procedure of EKF.

• From a theoretical point of view, the ideal candidate for FDI-applications seems to be the UIO. Here the principal problem (of quantitative model based approaches) to devide model uncertainties and faults is solved by a mathematical exact approach. The costs for the exact solution can not be payed for the application to LVS; the measurements are not available in practice.

• The same problem is connected to the application of the parameter identification approach (PIA) for advanced purposes. These techniques have their legitimation from

structural testing (of new structures etc.). For online-use during the life-time of the system the measurement effort is high or can not be fulfilled (criteria 5). The causal relation is principally not given by the parameter (a_i, b_i). To overcome these difficulties fast backtransformation techniques have to be developed to observe physical parameters.
• The application of the PIO is system-model dependent, but fault model independent, the measurement effort is low, the display parameters using the PIO-technique itself /21/ are mechanical states which has to be interpreted. To overcome these interpretation leakage with the goal to be closer to the physics of the fault, in /22/ a hypothesis testing strategy is suggested. Here the physical character of the fault can be displayed directly. Actual work is done to combine the testing strategy with pattern recognition qualities of neural networks /16/.
• As a provisional result it can be stated, that some research directions try to solve FDI problems related to the application to LVS. All approaches have advantages and disadvantages related to the interesting field. Actual developments try to overcome these difficulties. • To get a good base for the comparison of different techniques, a typical example should be formulated as a benchmark problem.

SUMMARY AND CONCLUSIONS: In this contribution the basic ideas of quantitative model - based fault detection and isolation are introduced and compared with practical proved and often implemented signal-based approaches. Concerning the interesting field of large vibrating structures some criteria are formulated for comparison of the techniques. As a result it can be stated that some techniques (Unknown Input Observer, Combined Parameter Identification - Extended Kalman Filter approach) are not problem adequate, other techniques (Parameter Identification Approaches, Signal-based approaches) can be applied, but do not solve the usefull causal relation in a adequate way. The remaining techniques to be considered (Proportional Integral Observer, Extended Kalman Filter, Parity Space Approach) can be applied, but have to be optimized to reach the causal relation: display parameter - physical fault. Further research work has to be done. For better comparison, a well-chosen problem adequate benchmark example should be chosen.

REFERENCES:

/1/ Basseville, M.; Nikiforov, I.V., 1993a, "Detection of Abrupt Changes", Prentice Hall, New Yersey.
/2/ Basseville, M.; Benveniste, A.; Gach-Devauchelle, B.; Goursat, M.; Bonnecase, D.; Dorey, P.; Prevosto, M.; Olagnon, M., 1993b, "In situ damage monitoring in vibration mechanics: diagnostics and predictive maintenance", *Mechanical Systems and Signal Processing*, vol. 7, no. 5, pp. 401 - 423.
/3/ Frank, P.M., 1990, "Fault diagnosis in dynamic systems using analytical and knowledge based redundancy - a survey and some new results", *Automatica*, 26, pp. 459-474.
/4/ Frank, P.M.; Köppen, B., 1993, "Review of optimal solutions to the robustness problem in observer-based fault detection", *Proc. Istn Mech Engrs, Part I: Journal of Systems and Control Engineering*, 207, pp. 105-112.
/5/ Frank, P.M., 1994, "Diagnoseverfahren in der Automatisierungstechnik", *at-automatisierungstechnik*, 42, pp. 47-64.
/6/ Fritzen, C.P., 1986, " Ein Beitrag zur Identifikation der Massen-, Dämpfungs- und Steifigkeitsmatrizen linearer mechanischer Systeme", Dr.-Ing. Thesis, University of Kai-

serslautern, Germany.
/7/ Fritzen, C.P.; Seibold, S., 1990, "Identification of Mechanical Systems by Means of the Extended Kalman Filter". *Proc. 3rd Int. IFToMM Conf. Rotordyn.*, Lyon, pp. 423-429.
/8/ Gertler, J., 1991, "Analytical redundancy methods in fault detection and isolation - a survey", *IFAC-Symposium Safeprocess*, Sept. 10-13, 1991, Baden-Baden, Germany.
/9/ Hou, M.; Müller, P.C., 1994, "Disturbance decoupled observer - a unified viewpoint", *IEEE Trans. Aut. Contr.*, AC-39, pp. 1388-1392.
/10/ Hou, M., 1995, "Descriptor Systems: Observers and Fault Diagnosis.", VDI - Fortschrittberichte, Reihe 8, Nr. 482, Düsseldorf, Germany.
/11/ Ljung, L., 1979, "Asymptotic Behaviour of the Extended Kalman Filter as a Parameter Estimator for Linear Systems", *IEEE Trans. Aut. Contr.*, AC-24, pp.36-50.
/12/ Patton, R.J.; Kangethe, S.M., 1989, "Robust fault diagnosis using eigenstructure assignment of observers", in: Patton, Frank, Clark, 1989, pp. 99-154.
/13/ Patton, R.J.; Frank, P.M.; Clark, R.N. (eds.), 1989, "Fault diagnosis in dynamic systems - theory and applications", Prentice Hall, Ney Jersey.
/14/ Pouliezos, A.D.; Stavrakakis, G.S., 1994, "Real Time Fault Monitoring of Industrial Processes", Kluwer, Dordrecht.
/15/ Rückwald, R., 1993, "Modellgestützte Überwachung mechanischer Systeme mittels differentieller Parameteridentifikation" VDI-Fortschrittberichte, Reihe 8, Nr. 328, VDI-Verlag, Düsseldorf, Germany.
/16/ Schwarze, H., 1995, "Fault detection of a mechanical system using a neural network", Student thesis, SRM, University of Wuppertal.
/17/ Seibold, S; Fritzen, C.P., Leifeld, A., 1993a, "A Combined State and Parameter Estimator Applied to Fault Detection", *IUTAM Symp. Ident. Mech. Syst.*, University of Wuppertal, Springer Verlag, appears.
/18/ Seibold, S.; Söffker, D.; Fritzen, C.P., 1993b, "Modellgestützte Detektion von Wellenrissen", *4th Meeting Dynamical Problems - Modeling and reality.*, eds: Natke, H.G.; Tönshoff, H.K.; Meltzer, G., University of Hannover.
/19/ Seibold, S., 1995a, "Ein Beitrag zur modellgestützten Schadendiagnose bei rotierenden Maschinen", VDI - Verlag, Düsseldorf, Germany.
/20/ Seibold, S.; Fritzen, C.P., 1996, "Identification Procedures as Tools for Fault Diagnosis of Rotating Machinery", *Int. J. of Rotating Machinery*, appears.
/21/ Söffker, D; Yu, T.; Müller, P.C., 1995a, "State Estimation of Dynamical Systems with Nonlinearities by using Proportional - Integral Observer", *Int. J. Syst. Science*, 26, no. 9, pp. 1571-1582.
/22/ Söffker, D., 1995b, "A new model-based tool for fault detection and isolation in machine- and rotordynamics", Proc. 11th Conf. Reliabiliy, Stress Analysis and Failure Prevention, Boston, Sept. 1995.
/23/ Wünnenberg, J., 1990, " Observer-based fault detection in dynamic systems", VDI Fortschritt Berichte, Reihe 8, Nr. 222, VDI-Verlag, Düsseldorf, Germany.

BEARING FAULT DETECTION VIA HIGH FREQUENCY RESONANCE TECHNIQUE WITH ADAPTIVE LINE ENHANCER

Li, Y., Shiroishi, J., Danyluk, S., Kurfess, T., Liang, S. Y.
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Abstract: This paper presents a new scheme for effective detection of rolling element bearing faults by the utilization of accelerometer signal processed by high frequency resonance technique (HFRT) with adaptive line enhancer (ALE). An envelope signal obtained by the HFRT can be considered as the sum of narrow band signals spaced at a characteristic defect frequency but corrupted by broad band noise. The ALE can be used to detect the presence of the narrow band components by separating narrow band from broad band signals. The HFRT takes advantage of the large amplitude of a defect signal over a range of high resonant frequency, then provides an envelope signal with a higher signal to noise ratio in the absence of low frequency mechanical noise. The ALE increases the detectability of the periodic bearing defect signal by providing an enhanced envelope spectrum with clear peaks at the harmonics of a characteristic defect frequency. The integration of the HFRT and ALE provides a useful tool in the early defect detection of rolling element bearings. Experiments for damaged bearings with defects have been performed. Results show that the scheme has substantially greater sensitivity than the HFRT without the ALE.

Key Words: Adaptive line enhancer; bearing; diagnostics; fault detection; high frequency resonance technique; vibration analysis.

Notations:

$\longleftrightarrow$:	Fourier transformation pair
*:	convolution operation
$s(t)$:	bearing defect signal
f_r:	resonant frequency
f_d:	characteristic defect frequency
$G(f)$:	as defined in equation (3)
$\mathcal{F}$:	Fourier transformation
$f_{\min}, f_{\max}$:	cutoff frequency of bandpass filter
$s(t), S(f)$:	bearing defect signal and its Fourier transform
$s_b(t), S_b(f)$:	bandpassed signal and its Fourier transform
$s_d(t), S_d(f)$:	demodulated signal and its Fourier transform

$s_e(t), S_e(f)$:	envelope signal and its Fourier transform
f_s:	sampling frequency
$x(k)$:	as defined in equation (7)
$n(k)$:	broad band noise
L:	filter length
Δ:	number of delay
X_k	input vector of filter
W_k	weight vector of filter
$y(k)$:	output of adaptive filter
$e(k)$:	error sequence
μ:	adaptive step size
P:	signal autocorrelation vector
R:	autocorrelation matrix of input signal
E:	expectation operation

INTRODUCTION: The initial fault detection of rolling element bearings of a rotating machine is important to the machine maintenance and automation process. Bearing maintenance based on statistics has many drawbacks. For example, it is possible to replace a bearing that has additional several hundred hours life by a new one with only several hours life. Unexpected foreign objects induced through lubrication system can dramatically shorten a bearing life. Therefore the monitoring and initial fault detection of rolling element bearings have received considerable attention for many years.

The generation of vibration due to a bearing defect was well explained by [1], [2] and [3]. Periodic impulses generate when a defect of a bearing contacts with another surface shown in Figure 1(a). The impulse includes a sharp rising edge and decays approximately exponentially due to system damping [2]. The short time duration of the impulse distributes its energy over a wide frequency band. Therefore, it is easily camouflaged by noise. But the impulses can excite a resonance at a higher frequency than the vibration generated by other machine elements with its energy concentrated in a narrow band around the resonant frequency that is relatively easy to detect. The frequency associated with the periodicity of the impulse is referred to as the characteristic defect frequency. It depends on the location of a defect, rotation speed and bearing dimensions, and can be calculated deterministically [4].

To detect bearing defects, many methods have been developed and studied. Time domain methods are usually based on the statistically different behaviors between good bearings and defective ones such as shock pulse counting [5], r.m.s., peak values, crest factor and kurtosis method ([6],[7],[8]). Some sophisticated pattern recognition methods [9] were developed to separate good from defect bearings. Before these methods are used, vibration signals are bandpassed at a high frequency range to reduce low frequency noise.

In the frequency domain, simple Fourier transformation at a low frequency range is limited with respect to bearing defect detection. The spectrum of a defect signal is spread across a wide frequency band and has very low amplitudes at the lower harmonics of its characteristic defect frequency as illustrated in Figure 1(a), while strong mechanical noise is usually concentrated in a low frequency band [10]. Therefore, new methods that are based on the statistic dependence of the harmonics of a characteristic defect frequency were developed. The signal average method [1] provides the total energy of the family of the harmonics of a characteristic defect frequency while the bicoherence analysis [11] reveals the statistics dependence of the phase of them. The high frequency resonance technique (HFRT) ([2], [10]) takes advantage of the large amplitudes of a defect signal in the range of a higher resonant frequency where low frequency noise is absent. Demodulation of the resonance makes it possible not only to detect the presence of a defect, but also to diagnose the location of the defect.

Most difficulties posed in bearing initial fault detection stem from the presence of a variety of noises and the wide spectrum of a bearing defect signal. Therefore, the success of bearing fault detection methods usually depends on how to increase a bearing defect signal to noise ratio. Dual-sensor based adaptive noise cancellation method [12] cancels only background noise and other noise originating from background-independent sources can not be treated. More than often, the signal of a damaged bearing is too subtle to be detected at the early stage of damage since it is hidden by noise. Although signal averaging [1] provides a method to reduce white noise, it strongly depends on the prior knowledge of a characteristic defect frequency to achieve a computationally efficient real-time analysis. It is also influenced by the noise having the frequency of some of the harmonics of the characteristic defect frequency.

In this paper, a noise cancellation based technique is proposed for bearing incipient defect detection. In this method, the adaptive line enhancer (ALE) is used to increase the detectability of a periodic defect signal. It enhances the spectrum of its envelope signal obtained by the high frequency resonance technique. In the following, the principle and application of the high frequency resonance technique by using a full-wave rectifier are explained first. Then the adaptive line enhancer with a recursive least mean square algorithm for wide band noise cancellation is presented.

PROPOSED SCHEME OF BEARING INITIAL FAULT DETECTION: The proposed detection scheme can be separated into two main parts: the first is the conventional high frequency resonance technique and the second is the adaptive line enhancer. The high frequency resonance technique takes advantage of the large amplitudes of a defect signal in the range of a high resonant frequency, and provides an envelope signal with a high defect signal to noise ratio in the absence of low frequency mechanical noise. The adaptive line enhancer can further reduce wide band noise of the obtained envelope signal and, therefore, enhance the envelope spectrum of the defect signal with clear peaks at the harmonics of the characteristic defect frequency. The whole scheme of bearing incipient fault detection is shown in Figure 2.

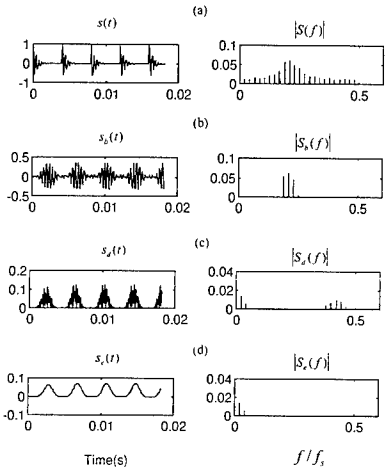


Figure 1. Process of the HFRT

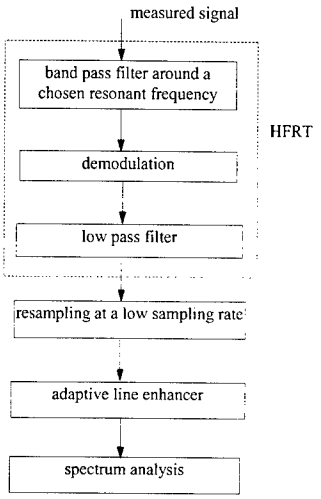


Figure 2. Diagram of proposed bearing initial fault detection

High frequency resonance technique: The high frequency resonance technique ([2], [10]) is illustrated in Figure 1 and involves 3 steps:

1. bandpass a measured signal around a selected high frequency band with the center at a chosen resonant frequency of a system.
2. demodulate the bandpassed signal by a non-linear rectifier.
3. use a low pass filter to cancel high frequency components and retain the low frequency information associated with bearing defects.

Based on the description of the signature of a bearing defect signal, the bearing defect signal can be denoted as [1]:

$$s(t) = e^{-\xi t}\cos(2\pi f_r t) * \sum_i \delta(t - i/f_d) \tag{1}$$

which is the bursts of exponentially decaying sinusoidal vibration at a system resonant frequency f_r. Its Fourier transform can be expressed as:

$$S(f) = G(f)\delta(f - if_d) \tag{2}$$

where

$$G(f) = f_d \mathcal{F}\left\{e^{-\xi t}\cos(2\pi f_r t)\right\} \tag{3}$$

After being bandpassed, the corresponding bandpassed signal $s_b(t)$ will be:

$$s_b(t) \longleftrightarrow S_b(f) = G(f)\delta(f - if_d) \quad where \quad f_{\min} < f < f_{\max} \tag{4}$$

The bandpassed signal is rectified by the following full wave rectifier:

$$s_d(t) = s_b(t)s_b^*(t) \tag{5}$$

Since a measured bearing signal is real, we have the following Fourier transform pair:

$$s_d(t) \longleftrightarrow S_d(f) = S_b(f) * S_b(f) \tag{6}$$

For any bandpassed signal $s_b(t)$ with the passband of $[f_{\min}, f_{\max}]$, the nonzero magnitudes of signal $s_d(t)$ only exist in the range of $\left[0,(f_{\max} - f_{\min})\right]$ and $2[f_{\min}, f_{\max}]$. In order to cancel the high frequency components and retain the low frequency information associated with a bearing defect, $s_d(t)$ is passed through a low pass filter. The output signal of the low pass filter is termed the envelope signal as denoted by $s_e(t)$. For a pure bearing defect signal $s(t)$, its magnitude is nonzero only at the harmonics of a characteristic defect frequency. Therefore, the spectrum of $s_e(t)$ will have nonzero values only at the harmonics of the characteristic defect frequency that are less than the bandwidth of a used bandpass filter shown in Figure 1(d). In theory, the bandwidth of a passband filter should be greater than the maximum characteristic defect frequency of a bearing in order to detect all possible defects. The nonzero values of an envelope spectrum at the harmonics of a particular characteristic defect frequency indicate the occurrence of a defect, and the location of the defect can be determined by its unique characteristic defect frequency.

Adaptive line enhancer [13]: In reality, measured vibration signals often include a variety of noises. The envelope signal of a damaged bearing obtained by the above method is contaminated by broad band noise making it difficult to detect the early damage of the bearing. From the above analysis, we see that a real time envelope signal $x(k)$ can be considered as the sum of narrow band signals $s_e(k)$ spaced at the harmonics of a characteristic defect frequency but corrupted by broad band noise $n(k)$:

$$x(k) = s_e(k) + n(k) \tag{7}$$

Therefore the adaptive line enhancer (ALE) can be used to separate the narrow band signals from the broad band noise. The ALE was first introduced by [13] and its performance was studied by [14]. The delayed version of $x(k)$ is used as the reference

input of the adaptive filter. The output of the filter is then subtracted from the primary input signal $x(k)$ to form an error sequence $e(k)$. The error sequence is fed back to adjust the filter weights shown in Figure 3.

The principle of the adaptive line enhancer can be understood as follows. The delay of an input signal causes the decorrelation between the broad band noise components of reference input and primary input while the narrow band signal is still highly correlated. To minimize the error power of the filter, the adaptive filter compensates for the correlated signal so that it can be canceled at the summing junction. Since the filter can not compensate for the decorrelation of the broad band noise components, only the narrow band signal is output from the adaptive filter. In the frequency domain, the filter weights will tend to form a bandpassed function about the center frequencies of the narrow band input components. Therefore, the broad band noise components of delayed input are rejected, while the narrow band signal information is retained. Here a recursive least mean squares (LMS) algorithm is used for this adaptive filter because of its computation simplicity. The output of the filter is:

$$y(k) = \mathrm{W}_k^T \mathrm{X}_k \tag{8}$$

where W_k^T is the weight vector of the filter and input vector X_k is

$$\mathrm{X}_k = \left[x(k-\Delta) \quad x(k-1-\Delta) \quad \cdots \quad x(k-L+1+\Delta)\right]^T \tag{9}$$

Then the error is:

$$e(k) = x(k) - y(k) \tag{10}$$

The mean square error (MSE) is the expected value of $e(k)$. LMS algorithm controls the weight vector of the filter so that the mean square error is minimized. The optimal weight vector W^* to minimize MSE is [13]

$$\mathrm{W}^* = R^{-1}P \tag{11}$$

where the input signal autocorrelation vector P is

$$P = E\left\{x(k)\mathrm{X}_k^{\mathrm{T}}\right\} \tag{12}$$

and the autocorrelation matrix R is:

$$R = E\left\{\mathrm{X}_k\mathrm{X}_k^{\mathrm{T}}\right\} \tag{13}$$

In practice, R and P are generally unavailable because they require an expectation operation. The following recursive LMS algorithm is a practical method to find a approximate solution of equation (11). The weights of the filter are updated by:

$$W_{k+1} = W_k + 2\mu e(k)X_k \tag{14}$$

where μ is adaptive step size. The simulation results in [15] show that it is appropriate to let μ satisfy the following stability condition in practice:

$$0 < \mu < \frac{1}{L\{power \quad of \quad input\}} \tag{15}$$

Therefore, given the input sequence $x(k)$ and initial value of the weight vector, the values of filter output $y(k)$, the error $e(k)$ and the weight vector W_k can be computed for all time by equations (8), (10) and (14).

Without the requirement of a priori information, the ALE's self-tuning capability allows it to find weak narrow band signals in wide band noise. Therefore, it has the potential to detect the line spectrum of the envelope signal of a defect bearing in a noisy environment.

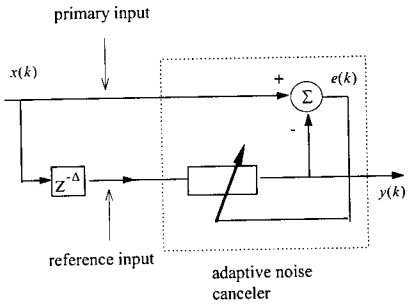


Figure 3. Block diagram of ALE

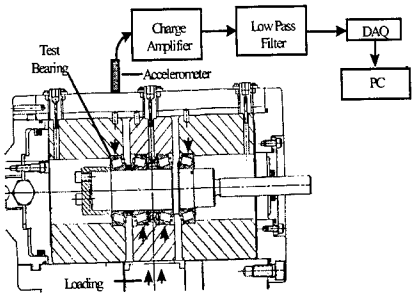


Figure 4. Test system schematic

EXPERIMENT AND DISCUSSION: Experiments have been performed to verify the proposed scheme. The experimental apparatus is composed of a base, test housing system, a drive system, and a data acquisition system. The test housing seen in Figure 4 has a 5 in. bore and a built in radial load cylinder. The radial load is given by a hydraulic pump. A Pacific Scientific DC servomotor along with a controller and flexible coupling is used to drive the shaft. The controller is programmed via serial link to a PC. The data acquisition system is a combination of a Pentium computer along with a National

Instruments DAQ-1200 PCMCIA data acquisition card. A Wilcoxan 736T high frequency accelerometer provides a voltage signal to the PC after it is fed through a Krohn-Hite programmable filter.

The test bearings are Timken LM501310 cup and LM501349 cone roller bearings. The defects were artificially made by scratching with a diamond scribe axially in the center of the outer race, and were positioned in the center of the loaded region. The bearing referred to as light defect bearing has a scratch on the order of 0.003 in. wide and 0.14 in. long. The heavy defect bearing has a larger scratch on the order of 0.006 in. wide and 0.12 in. long. Experiments with the good, the light defect, and the heavy defect bearings were performed under the radial load of 1100 lb. per bearing at 1200 rpm. The position of the test bearings and the load distribution can be seen in Figure 4. This type of defect in the bearings running at 1200 rpm results in a characteristic defect frequency of 164 Hz.

The used sampling rate was 20 kHz and the signals were prefiltered by a low pass filter with cutoff frequency of 10 kHz for anti-aliasing. It was found in the experiments that the mode of this system at the frequency of 8.2 kHz is sensitive to a bearing defect. Therefore, a digital bandpass filter with the passband of 7.8-8.6 kHz was used in the high frequency resonance analysis. After being demodulated, the signals were passed through a low pass filter with the cutoff frequency of 600 Hz. The obtained envelope signals were resampled at a sampling rate 20/15 kHz. The reason for resampling an envelope signal at a lower sampling rate is that its spectrum is less than the cutoff frequency of the used low pass filter. In addition, the decrease of sampling rate can increase the frequency discrimination of the ALE filter for a fixed filter length. A total of 32768 sampling points were obtained. After the envelope signals of 32768 points were resampled, 2048 points were used to train the ALE. The same data were filtered by the trained ALE for spectrum analysis. An ALE filter length of 100 and an adaptation step size of 10^{-6} were used.

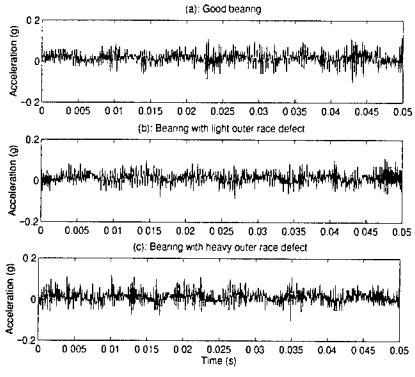


Figure 5. Raw vibration signals

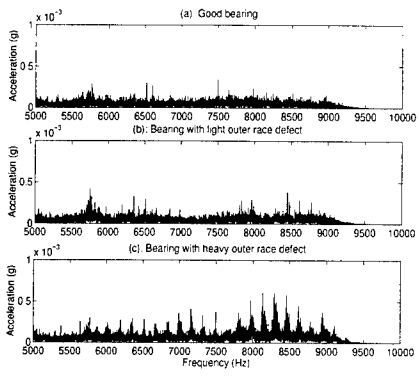


Figure 6. Spectrum of raw vibration signals

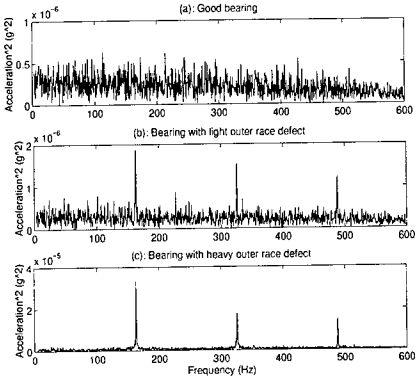


Figure 7. Envelope spectrum without the ALE

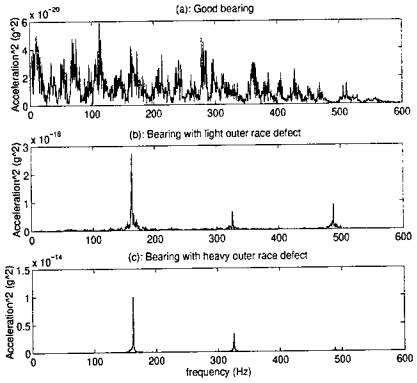


Figure 8. Envelope spectrum with the ALE

Experimental results are shown in Figures 5-8. There is no clear visible indication of a defect from the raw signals. The peaks spaced at the characteristic defect frequency are only shown around the frequency of 8.2 kHz for the bearing with heavy outer race defect in Figure 6. However, clear indications in terms of peaks at the characteristic defect frequency of 164 Hz and its harmonics for both defect cases can be highlighted by the further signal processing of the HFRT shown in Figure 7. It is evident that the ALE can significantly increase the detectability of the bearing defect signals by increasing the signal to noise ratios shown in Figure 8. As a quantifiable measure of bearing defect signals, the peak ratio is defined herein as the magnitude of an envelope signal at the characteristic defect frequency over the average magnitude of the envelope spectrum. Table I provides the peak ratios at the characteristic defect frequency of 164 Hz in the experiment.

Table I. Peak ratio with respect to defect size

	Heavy Damage	Light Damage	No Damage
Peak Ratio with ALE	173.45	66.70	4.34
Peak Ratio without ALE	36.76	7.22	2.31

CONCLUSIONS: In this paper, the scheme for early and effective defect detection of rolling element bearings was developed by the combination of the HFRT and the ALE. In this scheme, the HFRT was utilized to take advantage of the large amplitude of a defect signal over a range of high resonant frequency and the ALE was utilized to increase the detectability of the periodic bearing defect signal by separating it from the broad band noise. The experiments for damaged bearings with defects of different size have been performed. It is concluded that the ALE can significantly increase the

detectability of bearing defect signals and the combination of the HFRT and the ALE provides a sensitive and consistent method for detecting the early bearing defects.

The authors appreciate the Office of Naval Research for support through research grant N00014-95-1-0539, entitled "Integrated Diagnostics." Dr. Peter Schmidt serves as Scientific Officer. Content does not necessarily reflect the position or policy of the government, and no official endorsement is inferred. The Timken Company's assistance in the experimental set-up is also appreciated.

REFERENCE:

[1] Braun, S. and Datner B., 1979, "Analysis of Roller / Ball Vibration," *J. of Mechanical Design Transactions,* Vol. 101, pp. 118-125.

[2] Mcfadden, P. D. and Smith, J. D., 1984, "Vibration Monitoring of Rolling Element Bearings by the High Frequency Resonance Technique-A Review," *Tribology International*, Vol. 17, pp. 1-18.

[3] Li, C. James and Ma, Jun, 1992, "Bearing Localized Defect Detection Through Wavelet Decomposition of Vibrations," *PED-Vol. 55, Sensors and Signal Processing for Manufacturing, ASME,* pp. 187-196.

[4] Harris T. A., 1991, *Rolling Bearing Analysis*, pp. 950-951. John Wiley & Sons.

[5] Gustafsson, Olof G. and Tallian, Trbor, 1962, "Detection of Damage of Assembled Rolling Element Bearings," *ASLE Transactions*, Vol. 5, pp. 197-209.

[6] Dyer, D., and Stewart, R. M., 1978, "Detection of Rolling Element Bearing Damage by Statistical Vibration Analysis," *J. of Mechanical Design Transactions,* Vol. 100, pp. 229-235.

[7] Alfredson, R. J., and Marhew, J., 1985, "Time Domain Methods for Monitoring the Condition of Rolling Element Bearings," *Mechanical Engineering Transactions*, pp. 102-107.

[8] Martin, H. R. and Honarvar, F., 1995, "Application of Statistical Moments to Bearing Failure Detection," *Applied Acoustics*, Vol. 44, pp. 67-77.

[9] Li, C. James and Wu, S. M., 1989, "On-line Detection of Localized Defects in Bearing by Pattern Recognition Analysis," *J. of Engineering for Industry,* Vol. 111, pp. 331-336.

[10] Su, Y. -T. and Lin S.-J.,1992, "On Initial Fault Detection of a Tapered Rolling Bearing: Frequency Domain Analysis," *J. of Sound and Vibration*, Vol. 155, pp. 75-84.

[11] Li, C. James, Ma, J., and Hwang, B.,1995, "Bearing Localized Defect Detection by Bicoherence Analysis of Vibrations," *J. of Engineering for Industry*, Vol. 117, pp. 625-629.

[12] Carney, Matthew S., Mann, J. Adin III and Gagliardi, John, "Adaptive Filtering of Sound Pressure Signals for Monitoring Machinery in Noisy Environments," *Applied Acoustics,* 1994, pp. 333-351.

[13] Widrow, B. et al., "Adaptive Noise Canceling: Principles and Applications," *Proc. IEEE,* Vol. 63, 1975, pp. 1692-1716.

[14] Treichler, John R., "Transient and Convergent Behavior of the Adaptive Line Enhance," *IEEE Transaction on Acoustics, Speech and Signal Process,* Vol. 27, No. 1, 1979, pp. 53-62.

[15] Clarkson, Peter M, 1993, *Optimal and Adaptive Signal Processing*. pp. 166-173. CRC Press, Boca Raton, Florida.

EXPERIMENTAL AND NUMERICAL STUDY OF GEAR FAULTS IN GEARBOX: EVALUATION OF DYNAMIC STRESS FLUCTUATIONS

J.Mahfoudh, C.Bard, C.Beigneux and D.Play

Laboratoire de Conception et Analyse de Systèmes Mécaniques
Institut National des Sciences Appliquées de Lyon - Bat. 113
20, Avenue Albert Einstein - 69621 Villeurbanne Cedex - France

Abstract: Several studies were made on the effects of gear faults on vibrations and noise. Fault signatures were defined and procedures for fault detection were proposed. Now, the question is to define fault thresholds. As an experimental determination does not appear simple, thresholds can be defined by comparing the dynamic tooth root stress due to a fault to a fatigue limit of the gear material and by relating the type and the depth of the fault to the vibration levels and tooth root stress increase.
The experimental study of the gear fault effect on bearing vibrations enables us to define descriptors which indicate the presence of fault and its nature, the experimental study will be shortly summed up, then a numerical model will be developed involving validation with respect to previous experimental studies. The dynamic tooth load and tooth root stress with and without fault will be presented and the risks of tooth failure due to stress fluctuations were discussed.

Key Words: Condition monitoring, diagnostics, dynamic stress, gears, maintenance, numerical simulation, vibration analysis.

INTRODUCTION: This work takes place in a research program dealing with the preventive maintenance applied on gearboxes (Alattass, 1994). The aim is to define a maintenance policy which required many steps:
- observation of the mechanical behaviour of the studied system,
- detection of a change in this behaviour which indicates a modification of the mechanical characteristics of the system,
- determination of the origin and the nature of this change,
- finally, the definition of warning levels.

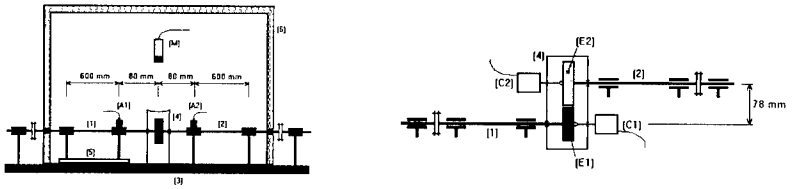
In order to achieve this task, experimentation were carried out, only mechanical behaviour modifications due to operating faults on tooth gears were considered. Vibrations from bearing housing seem to be a suitable system response to be used. The working method is as follow (Alattass et al, 1994 a):
- making tests with gears without fault,
- create fault on one pinion, then reproducing the same set of tests,
- processing the acceleration signals from bearing housing and analysing the signals in different domains (time, frequency and time-frequency (Wavelet)),

- extracting the influence of signal parameters through statistical design procedures
- establishing a methodology to identify the nature of fault, based on the use of statistical tests applied on the Wavelet analysis: KHI2 test identifies distributed fault made by scuffing, while "extreme values" test identifies local fault obtained by pitting (Alattass et al, 1994 b),
- definition of thresholds by comparing the maximum stress to a limit value. This step is carried out actually using a numerical model, the measurement of stress during function is still an expensive operation.

It is an evidence that experimental studies take lots of time, but one have to recognise that the experimental results must be magnified when theoretical or numerical models are used to understand trends and results. In the field of gear calculations during the last 10 years, fine dynamic gear behaviour were achieved through a precise tooth geometry definition and a kinematics transmission error determination. Consequently, numerical tools were available to simulate gear dynamic behaviours. The purpose of this paper is to define the relation between the presence of a tooth fault and the tooth failure through an increase of the fault tooth dynamic load.

EXPERIMENTAL PROCEDURE: The test stand (Fig. 1) is constituted of two shafts of 60 mm in diameter, mounted on two rolling bearings and coupled with gears. The input shaft (2) was driven by a 120 kW DC motor. The output shaft is braked by a DC motor. Rotating speed was varied between 0 and 6000 r/min and was feedback controlled. The input torque was varied independently from 0 to 150 Nm. The details of the test mountings were described elsewhere (Remod, 1991). Vertical and horizontal bearing accelerations have been measured with piezoelectric accelerometers A1 and A2 respectively. Other parameters (speed, torque, oil temperature) were also recorded for each test

The test gears were AGMA 12 quality with 36/38 teeth, a contact ratio of 1.3 and a pitch module of 2 mm. The chosen local fault simulated was pitch line pitting on a single tooth. The single line of pitting was modelled as a strip of metal removed at the pitch line. Three faults were created, their positions were determined using a marking compound (Fig 2). The geometrical characteristics of local faults for tested gear are gathered in the Table 1.



(M) microphone,(A1) and (A2) accelerometers, (C1) and (C2) optical encoders, (E1) and (E2) gears, (1) and (2) shafts, (3) rigid frame, (4) oil box, (5) misalignment plate, (6) isolating case

Fig. 1: Schematic arrangement of the test gear apparatus

The operating test conditions were : five speeds of rotation (1000, 2250, 3500, 4750 and 6000 r/min), and five applied loads (0, 30, 60, 90 and 120 Nm).

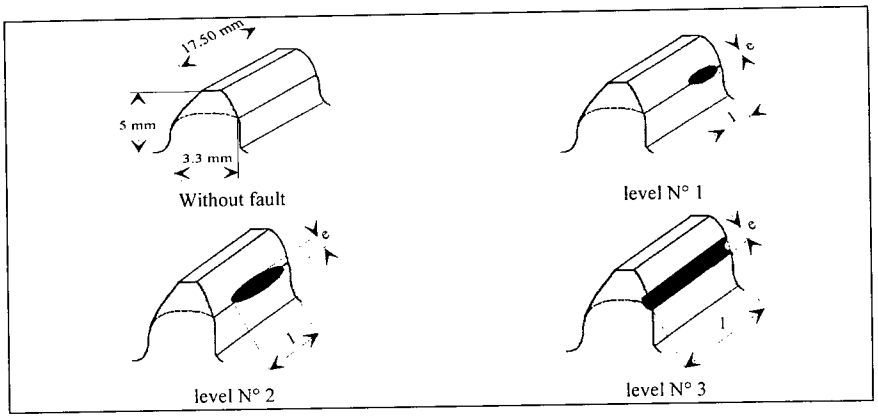


Fig. 2 : Visualisation of different levels of local fault

Table 1 Geometrical characteristics of local fault

	Width (l)	Depth (p)	thickness (e)
Level N° 1	4.4 mm	0.043 mm	0.4 mm
Level N° 2	8.7 mm	0.072 mm	0.75 mm
Level N° 3	17.5 mm	0.23 mm	0.97 mm

For each test, bearing accelerations were measured, then processed and analysed in time, frequency and Wavelet domains. The spectra was obtained from an average of 16 successive Fast Fourier Transforms of 1024 points signal acquisition. The acceleration Root Mean Square (r.m.s.) levels are compared with and without faults (no units for acceleration).

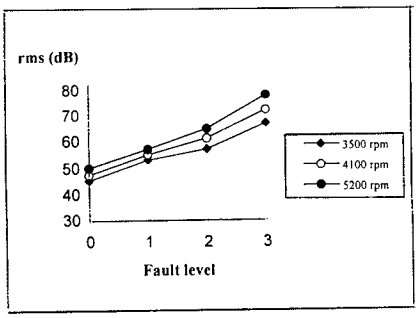


Fig 3 Effect of local faults on r.m.s. :Spectral analysis

Figure 3 shows that the bearing acceleration rms level increases when the size of the tooth fault increases in spectral domain, the same trends have been obtained in time and Wavelet analysis. In summary, a tooth fault can be detected, but now the question is to anticipate the tooth failure as soon as possible. Obviously, experimental determination can be dangerous when one large tooth fault gives a catastrophic tooth failure and even test rig failure. In the opposite side, a small tooth fault can give long endurance gear tests not always consistent with laboratory test conditions. Consequently, numerical tooth fault gear dynamic behaviour might be considered.

NUMERICAL APPROACH: A mathematical model has been developed to simulate tooth line pitting effect. It is based on the general equations of motion for a geared system (Blankenship and Singh, 1992). The experimental test rig was modelled using the Finite Elements Method, gears were represented with two rigid disks related by the meshing stiffness, the gear excitations are introduced by both the fluctuations of meshing rigidity and the static transmission error (Bard, 1995). These values were calculated by taking into account all the design parameters of the gears, using a software based on the Finite Prism Method (Olakorede and Play, 1991). This technique, used in the quasi-static analysis, allowed a 3D evaluation of the cylindrical gear elastic deformations with a fine definition of the load sharing and pressure distribution. A preliminary analysis begins by a geometrical description of the gear teeth, including profile modifications due to tooth fault, manufacturing errors, gear bodies position including eccentricity, centre distance and misalignments, then for each meshing position, potential areas of gear contact are located, and the unloaded transmission error and the relative distances between teeth are calculated and stored. Elastic deformations are then evaluated, local and global deformations are evaluated separately, and surface deformations are estimated by the well known Boussinescq theory. Final results give pressure distributions along the tooth face width, load sharing between different tooth pairs, stress, meshing stiffness and loaded transmission error. The general form of equations of motion is :

$$M.\ddot{X} + C.\dot{X} + K(\theta^*).X = K(\theta^*).\left\{\xi_o(\theta^*) + \varepsilon_o(\theta^*)\right\} \qquad (1)$$

where θ^* is a space parameter that describes the nominal progression of meshing. It could be expressed by $\theta^* = \Omega^*.t$, where Ω^* is the nominal angular velocity. $K(\theta^*)$ is the general form of the stiffness matrix (it depends on θ^* due to the meshing stiffness variations) and $\xi_o(\theta^*)+\varepsilon_o(\theta^*)$ is a specific vector which simulates excitation associated with the static and the unloaded transmission error (Özguven and Houser, 1988). M and C are mass and damping matrix.

The tests rig was modelled by 41 elements with 251 degrees of freedom. The chosen modal damping was 0.02 for all modes of the structure and 0.06 for gear mesh. The Pseudo-modal techniques were applied in order to reduce the number of equations, the reduced system of equations was solved using the well known step by step Newmark techniques.

The numerical model was validated with different mechanical cases in previous studies. For the quasi-static part, tooth root stress for cylindrical gears have been measured quasi-statically and the difference with the numerical simulation was less than 2% (Kim, 1996). For the dynamic part, good agreement was observed for the localisation of modes and critical speeds between the measured and the simulated dynamic transmission error (Bard, 1995).

Local faults were considered in the quasi-static analysis as a modification of the tooth geometry, the mesh stiffness has been calculated for the three levels noted above. (Fig. 4). It can be noted that modifications due to levels 1 and 2 are almost negligible and those due to level 3 reach a 50%.

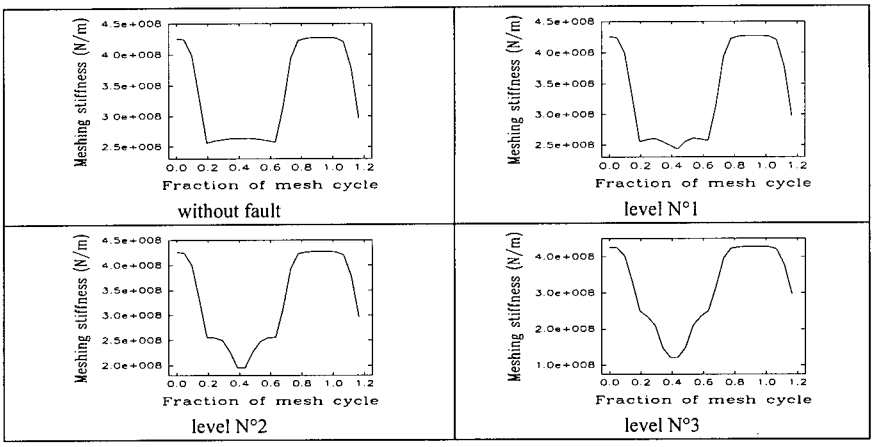


Fig. 4: Effects of faults on meshing stiffness evolution

Obviously, a geometrical profile variation due to gear fault leads to a modification of the gear kinematics motion. In other words, a pseudo-polygonal effect appears due to the flat bearing segment along the tooth profile caused by the fault geometry. The static transmission error was determined using the kinematics simulation of mesh for fault level 3 only (Fig. 5).

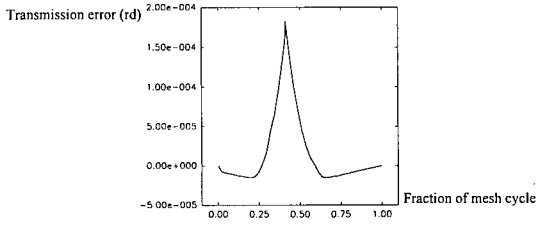


Fig. 5: Transmission error-fault level N°3

For the others levels, it was not possible to be simulated, there is always a contact area between teeth as the fault does not occupy the full tooth flank, that is why we have considered this kinematics excitation negligible, even if there is a modification of the measured transmission error.

Calculations were performed for different values of input torque . The maximum dynamic load is presented for the three levels of fault and for the maximum torque in figure 6. The increase due to fault level 3 reached 60%, while fault level 1 showed no effects and level 2 produced an increase slightly inferior to 10%.

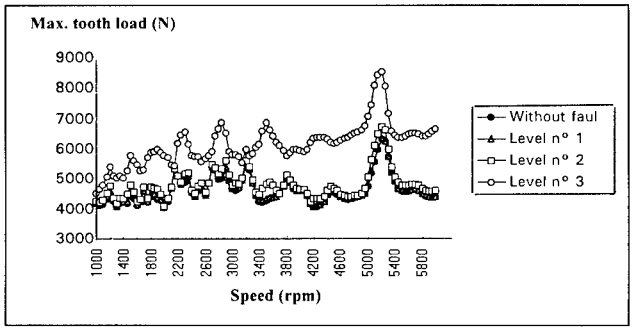


Fig. 6 : Maximum dynamic load versus speed

The evolution of the maximum dynamic load was calculated for the three levels at 4100 rpm and for different mesh positions (Fig. 7a), which represents one of the input data in the quasi-static analysis for stress evolution. The same trends have been observed for other speeds (Fig. 7b), we can deduce that the main increase is due to level 3 with the static transmission error, the fluctuations of the mesh rigidities have obviously no significant effects.

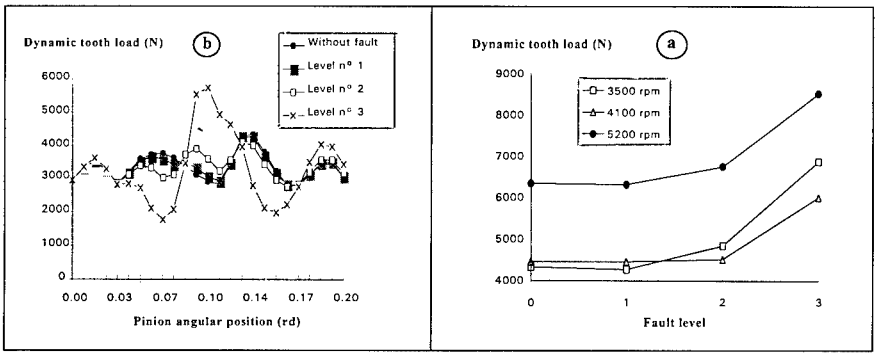


Fig. 7: Maximum dynamic load evolution

The tooth root stress distributions are compared with and without fault (Fig. 8). It can be noted that for level 3 the stress increase is located at the fault passage, for level 2 we can observe an increase due to fault passage in the opposite side of tooth width, and a fail of stress at fault position due to the loose of contact between teeth. Fault level 1 have no negative effect, this fault seems to have an effect of tooth correction

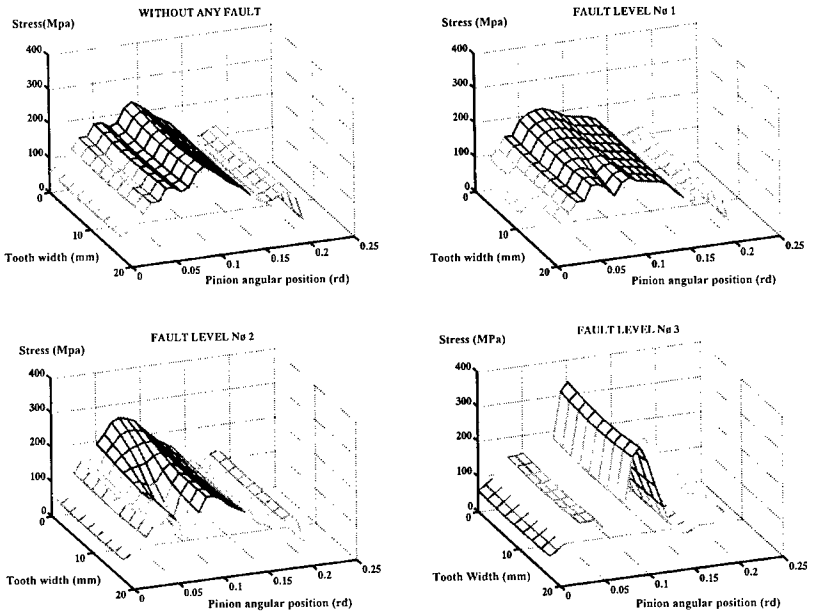


Fig. 8: Stress distributions

Stress values were compared in a cross section during tooth passage. It can be noted that level 1 and 2, as shown previously, have no significant effects (Fig. 9a), and the evolution of the stress for different mesh positions were compared in order to extract general trends relating fault geometry and tooth root stress (Fig. 9b).

The failure of stress level 3 for mesh position 0.08 and 0.15 rad is due to the looseness of contact between teeth, otherwise, the general trends are the same as for maximum load (Fig. 7b) and the measured acceleration levels (Fig. 3).

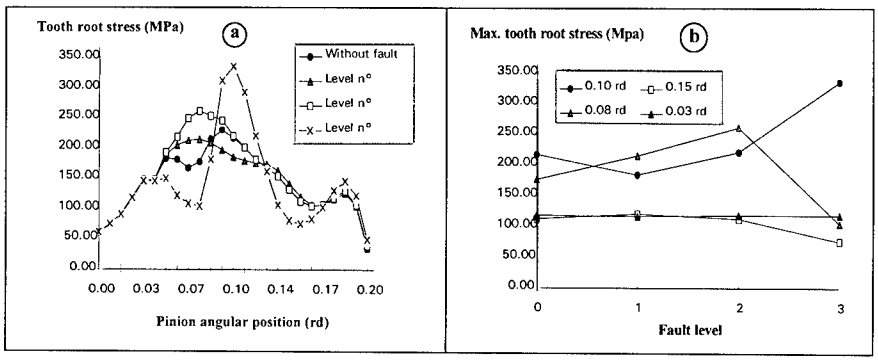


Fig. 9: Stress evolution

CONCLUSION: The aim of this work is to establish relationship between fault geometry, acceleration levels and the dynamic stress at tooth root, in order to define thresholds for gear local fault. The work was carried out using a numerical model, the faults were considered as a modification of the tooth flank, which induce a modification on the mesh rigidity and the static transmission error (level 3 only). The fluctuation of the dynamic load and stress distribution due to faults were calculated. The same trends have been observed for the measured acceleration and the simulated results. The modifications of the mesh rigidity have no significant effects on the dynamic behaviour of the studied system, excitation seems to be mainly influenced by the transmission error. It can be noted that the rate of increase for acceleration rms and the maximum dynamic load due to the fault evolution from level 2 to level 3 is almost the same (30%), those trends are similar to those of a failure rate curve for machinery (bath-tub curve) when the increase of the probability of failure after the component has exceeded normal useful life. Our aim actually is to have a real characterisation of the exciting forces by taking into account the measured transmission error as data for the numerical model, then results could be compared to crack growth models.

ACKNOWLEDGEMENTS: Many students in Master and PhD have participated to this work, the technical part have been accomplished by Soleilhac P., authors thank all this colleges

REFERENCES

ALATTASS, M., "Maintenance des machines tournantes: signature de defauts d'engrenage droits et helicoidaux," Thèse de doctorat, I.N.S.A de Lyon, 1994, 198 p.

ALATTASS, M., MAHFOUDH, J. and PLAY, D., "Vibration analysis for detection of gear faults within gearbox: application to spur and helical gears," Fourth

international conference on rotor dynamics, IFToMM, 1994, PP 197-205, Vibration Institut.

ALATTASS, M., MAHFOUDH, J. and PLAY, D., "Morphological study of vibratory signals," Conference on the acoustical and vibratory surveillance methods and diagnostic techniques, Paris, 1995, pp 456-474, CETIM.

BARD, C., "Modélisation du comportement dynamique des transmission par engrenages, "Thèse de Doctorat, INSA de LYON, 1995, 292 p.

BLANKENSHIP, G. W. and SINGH, R., "A comparative study of selected gear mesh interface dynamic models," International Power Transmission and Gearing Conference, A.S.M.E, 1992, Vol. 1 pp 137-146.

KIM, Y. C., "Distribution des charges et des contraintes dans les engrenages cylindriques avec voile. Simulation numérique par la méthode des Prismes Finis," Thèse de Doctorat, INSA de LYON, 1996, 151 p.

OLAKOREDE, A. and PLAY, D., "Load sharing load distribution and stress analysis of cylindrical gears by Finite Prism Method in a CAD environment," Proc. Design Productivity Int. Conf., Feb. 1991, Honolulu, Ragsdell & T.Holt, pp 921-927.

ÖZGUVEN, N. and HOUSER, D.R., "Mathematical models used in gear dynamics - a review," Journal of Sound and Vibration, 1988, N° 121(3) pp 383-411.

REMOND, D., "Contribution à l'étude et l'analyse expérimentale du bruit d'engrénement," Thèse de Doctorat, INSA de LYON, 1991, 231p.

FAULT DIAGNOSIS OF ROLLING ELEMENT BEARINGS USING ARTIFICIAL NEURAL NETWORKS

Lakshmi Rajamani
Department of Computer Science and Engineering
Osmania University
Hyderabad - 500 007, India

R. Dattagupta
Department of Computer Sciences and Engineering
Jadavpur University
Calcutta - 700 032, India.

Abstract : The basic concept of neural network and its application to diagnosis of rolling element bearings is described. Neural networks based on neurobiological systems are more adept at classification and identification tasks than conventional statistical and expert systems. In this paper the unsupevised learning method based on Kohonen network is used to study the classification of problem of common faults in rolling element bearings. Example patterns based on vibration analysis are used to train the network. The method successfully predicts the class of faults.

Key Words : Fault diagnosis, artificial neural network, rolling element bearings, vibration analysis, Kohonen network.

INTRODUCTION : Rolling element bearings are one of the most important single components in determining the reliability of a mechanical plant. Since these bearings are widely used in practically all rotating and reciporcating machines, it is essential to detect faults in the bearings. Rolling element bearing condition monitoring has received considerable attention because the majority of problems are caused by faulty bearings. The failure of these bearings is due to localized defects resulting in fatigue cracking [1]. An effective condition monitoring program must recognize the characteristics, conditions and developing trends of the operating mechanical system to predict and to make a decision for the future. To be truly effective a diagnostic system should be capable of indicating the locations, causes and severity of faults utilizing information derived from active operating conditions.

Rolling element bearings support load by means of a finite number of rolling elements and therefore produce vibration and sound. These vibrations arise from bearing element rotations, resonance of bearing elements and supports, intrusive vibrations (transmitted from shaft imbalance, gear meshing, vane passing etc.). Mechefske and Mathew [2] developed an unsupervised classification program 'SNOB' to automatically diagnose faults in low speed rolling element bearings. They however, observed that too large a learning set of spectra may prove to be undesirable when the variability of the spectra is significant. As artificial neural networks have high parallel information processing ability, self organization, self learning and much stronger associative memory, they have provided another way for fault diagnosis.

Unal [3] applied back propagation neural network for ball bearing monitoring and diagnosing. In this paper, the unsupervised learning method based on Kohonen network is used to study the classification problem of common faults in rolling element bearings.

ARTIFICIAL NEURAL NETWORKS : Artificial Neural networks provide a data based, heuristic approach to condition monitoring and diagnostics for rotating machinery. Neural computing systems are more adept at many classification identification tasks than both traditional statistical and expert systems. These tasks require the ability to match large amounts of information simultaneously and then generate categorical and generalized outputs. Neural networks have capacity to learn and store information from past operating faults via associative memory, Neural computation models are specified by three basic entities [4].

- node properties, the activation range and the activation function.
- the network architecture (topology), the types of connection, the nodes of connection and the weight range.
- system dynamics, the weight initialization scheme, the activation calculation format and the learning rule.

Topology : The topology of a network refers to its framework as well as its interconnection scheme. The framework is often specified by the number of layers and the number of nodes per layer. The types of layers include the input layer, a hidden layer and the output layer. According to the network connection a network can be either feed forward or recurrent and its connection either symmetrical or asymmetrical.

Learning : Learning is one of the most important features of artificial neural network. It is emphasized that there is no known method of finding learning rules. However, once a learning rule has been proposed, one can analyze and study its properties. The classification schemes that have been proposed are

A) Interaction with the Environment

1. Supervised learning
2. Reinforcement learning
3. Unsupervised learning

B) Topological changes

1. Weight change only
2. Topology change only
3. Weight and topology change.

C) Stochastic and deterministic learning

Kohonen Network : Kohonen network which is used to study the classification of bearing faults is an unsupervised learning method. Each node in the Kohonen network receives inputs from the environment and from other nodes in the layer. When we build a Kohonen network it is important to properly initialize the weight vectors of the nodes. Both input vectors and weight vectors are normalized to a constant. Each node computes the Euclidean distance by taking the dot product of its weight vector and the input vector. Kohonen network performs clustering through a competitive learning mechanism called 'winner take all'. The node with the largest activation level is declared the winner in the competition. This node is the only node that will generate an output signal, and all other nodes are suppressed to the zero activation level. Furthermore, this node and its neighbors are the only nodes permitted to learn for the current input pattern.

The Kohonen network uses intralayer connections to moderate this competition. The output of each node acts as an inhibitory input to the other nodes but is actually excitatory in its neighborhood. Thus, even though there is only one winner node, more than one node are allowed to change their weights. This complex scheme for moderating competition within a layer is known as lateral inhibition. The inhibitory effect of a node can also decrease with the distance from it and assume the appearance of a Mexican hat. The exact size of the neighborhood varies as the learning goes on. It starts large and is slowly reduced, making the range of change sharper and sharper. To simulate lateral inhibition, one simply takes the winner. The winner node and its neighbors will learn by adjusting their weight vectors. Kohonen learning is to make the winning weight vector more similar to the input vector. As learning proceeds, the size of the neighborhood is gradually decreased. Fewer and fewer nodes learn in each iteration and finally only the winning node learns.

The Kohonen algorithm which has been used is as follows :

Step 1 : Initialize weights

Initiatize weights from N inputs to M outputs using small random values. Set the initial radius of neighborhood and initial learning rate.

Step 2 : Present new input

Step 3 : Compute distances to all nodes. Compute distance dj between the input and each output node j using.

$$dj = \sum_{i=0}^{N-1} (Xi\ (t) - Wij\ (t))^2$$

where Xi(t) is the input to node i at time t and Wij(t) is the weight from input node i to output node j at time t.

Step 4 : Select output node with minimum distance. Select node j* as that output node with minimum distance dj.

Step 5 : Update weights to node j* and neighbors weights are updated for node j* and all nodes in the neighborhood. New weights are

$$Wij(t+1) = Wij(t) + \alpha(t)\,(Xi(t)-Wij(t))$$

The term α (t) is a gain term between 0 and 1 that decreases in time.

Step 6 : Repeat by going to step 2.

VIBRATION CONDITION MONITORING TECHNIQUES : An excellent description of monitoring vibration signatures from rolling element bearings, generation of fault related signal and vibration criteria based on experience is available [5, 6]. Several techniques which are available for finding defects in rolling element bearings rely on the generation of defect induced vibration within the bearing. During the bearing operation, wide band impulses result from the passage of defect through the rolling element, race way contacts. Defects at different locations of a bearing (inner race, roller, outer race) have characteristic frequencies [7] at which impulses are generated. The characteristic defect frequency makes it possible to detect the presence of a defect and diagnose in what part of the bearing the defect is present. The difficulty in localized defect detection, however lies in the fact that the energy of the bearing signature is spread across a wide frequency band and therefore some of the characteristic frequencies can be buried by noise (due to the fact that signals due to intrusive vibration have sufficient strength). This problem has been practically overcome by using time domain methods involving indices that are sensitive to the impulse oscillations. Examples are RMS level, peak level, crest factor (or impact index), Kurtosis value and shock pulse counting. These methods would yield satisfactory results as long as one selects a frequency band that includes some of the modes that get actually excited by impact of rollers and where there are no other sources of impact. But unfortunately, it is difficult to predict vibration modes of bearing's parts and its supporting structure, and which one of these modes is likely to be excited by the impact on the defect. A mode excited at one stage of a defect may not be excited when the extent of defect increases.

A number of frequency domain techniques have been developed to detect a train of impulses having some characteristic defect frequency in the bearing signal. Among these techniques are synchronous averaging, cepstrum analysis, high frequency resonance, sum and difference frequency analysis. Barkov and Barkova [6] found the high frequency vibration envelope method to be the most efficient for rolling element bearings condition assessment and with proper measurement the assurance for life

time prediction. It is assumed that the defect frequencies with corresponding amplitudes have been obtained

IMPLEMENTATION AND RESULTS : The computer program is developed in C programming language with Borland C/C++ compiler. The program works for any number of input nodes and any number of output classes. The patterns for training are generated from data available in literature [7]. The input vector can have fourteen components. The first component is the first harmonic component of speed, the second element is the second harmonic component of speed, the other elements are vibration amplitudes corresponding to the roller, outer and inner race frequencies and its harmonics (upto 4th harmonic). The above input is classified into 6 classes. This data is fed into the program. The output layer may be 1-dimensional or 2-dimensional. In the present study, 2-dimensional output layer is considered. The output layer size is not known initially. Therefore, it is set to number of output classes. The memory is allocated for input vector, the number of random numbers required are determined from number of input nodes and number of output nodes. For a 2-d network, the number of random numbers required for weight initialization is

$$\text{rand} = m \times n \times p$$

n is the number of inputs, m is the number of rows and p is the number of columns in the network. The network is tried with five elements which correspond to the first and second harmonic components of speed and the first harmonic of the characteristic defect frequencies. Input faults are generated. Fifty example patterns are used to train the network. For each example pattern used for training the network, euclidean distances are computed, the minimum distance determined and thus the winner node for that example is found out. This training is done for all cycles. If the neighborhood is '0' or the learning rate is '0', then the loop breaks and the next part of the program is executed where the number of output classes are determined. Otherwise, at the end of fifth cycle, the number of classes obtained for the given input examples are determined. If this number is equal to the number of output classes the data should be classified into, then the network weights are finalized; if not the output layer size is incremented and all steps mentioned above are repeated until the data is classified. The final weights and the network dimensions are copied into a data file and the program for training terminates. The given input samples are classified into unbalance, misalignment, roller defect, inner race defect, outer race defect, no defect.

The input vector has five components. With 1-dimensional output layer convergence was not achieved and therefore 2-dimensional output layer was chosen. Network output layer size for the given input data fixed by trial and error is 13 x 13. The neighborhood value used in the present problem is 5 and is application dependent. The learning rate at the beginning of training is 0.65 and is reduced by 0.1 after each cycle. The program works through all the steps of Kohonen algorithm and calculates the number of classes in the output layer. The network is said to converge, when the weights of the network remain practically the same for two consecutive cycles.

Testing patterns used for training the network are given below :

Sl.No.	Testing Pattern				
1.	1	2	0.375	0.37	0.8
2.	2	2	0.374	0.371	0.8
3.	1	3.28	0.375	0.37	0.8
4.	1	2	0.9	0.37	0.8
.....					
49	1.17	2.1	0.42	0.42	1.01
50	1.1	2.09	0.41	0.4	2.6

The defects and corresponding nodes after training are given below :

Sl. No.	Type of Defect	Node activated
1.	Unbalance	[8,1]
2.	Misalignment	[8,5]
3.	Roller defect	[9,0], [10,12]
4.	Inner race defect	[5,6], [1,9]
5.	Outer race defect	[2,3]
6.	No defect	[5,4]

The patterns after training are as follows :

Sl. No.	Training Pattern					Winner Node
1.	1.0	2.0	0.375	0.370	0.80	[5,4]
2.	2.0	2.0	0.374	0.371	0.80	[8,1]
3.	1.0	3.28	0.375	0.370	0.80	[8,5]
.....						
49.	1.17	2.10	0.420	0.420	1.01	[5,4]
50.	1.10	2.09	0.410	0.400	2.60	[11,8]

Testing module execution gives results as follows :

Enter Input pattern : 1.150 2.010 0.370 0.380 0.850
winner node is : [5,4]
no defect

Try another input pattern ? If yes enter 1 : If no enter 0:
1
enter input pattern : 1.350 2.110 0.900 0.390 0.890
winner node is : [10,12]
roller defect

Try another input pattern ? If yes enter 1 : If no enter 0:
1
enter input pattern : 1.350 3.300 0.390 0.410 0.910
winner node is : [8,5]
misalignment

Try another input pattern ? If yes enter 1 : If no enter 0:
0

CONCLUSIONS : The artificial neural network is successfully tried on a typical bearing and the Kohonen method of unsupervised learning predicts the class of fault. Therefore the artificial neural network can be effectively used with an online monitoring system with spectral analysis capability to predict faults.

REFERENCES :

(1) Widner, R.L. and Littmann, W.E., "Bearing Damage Analysis", National Bureau of Standard special publication No.423, April 1976.

(2) Mechefske, C.K. and Mathew, J., "Diagnosing Faults in Rolling Element Bearings using Parametric Frequency spectra and 'SNOB' unsupervised classification". Proceeding of the 3rd International Machinery Monitoring and Diagnostics Conference, Las Vegas, 1991, pp 237-242.

(3) Unal, Aynur, "Detecting Ball Bearing Defects using Artificial Neural Networks", J. Acoustical Soc. America, 1 (88), Nov. 1990, p 520.

(4) Fu, L., "Neural Networks in Computer Intelligence", International Edition, Mcgraw Hill, 1994.

(5) Bannister, R.H., "A Review of Rolling Element Bearing Monitoring Techniques", Condition Monitoring of Machinery and Plant, published by Mech. Engg. Publications Ltd. for I.Mech. E. London, 1985, pp 11-24.

(6) Barkov, A.V. and Barkova, N.A., "Condition Assessment and Life Prediction of Rolling Element Bearings - Part 1", Sound and Vibration, Vol.29, No.6, 1995, pp 10-17.

(7) Bruel and Kjaer, "Machine Health Monitoring using Vibration Analysis", Lecture notes prepared for a course organised by Canadian Acoustical Association, BA 7095-13, 1983.

APPENDIX

MFPG/ MFPT PUBLICATIONS

Both printed and microfiche copies of the following MFPG publications whose catalog numbers start with either **AD**, **COM** or **PB** may be obtained from the

National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161

Glossary of Terms AD 721 354

MFPG Proceedings

Meeting Nos. 1-9 (set of five) AD 721 359

Meeting Nos. 1-5
Papers and Discussion on Failure Analysis and Control

Meeting No 6 "Detection, Diagnosis and Prognosis"
December 6, 1968

Meeting No 7 "Failure Mechanisms as Identified with Helicopter Transmissions"
March 27, 1969

Meeting No 8 "Critical Failure Problem Areas in the Aircraft Gas Turbine Engine"
June 25-26, 1969

Meeting No 9 "Potential for Reduction of Mechanical FailureThrough Design Methodology"
November 5-6, 1969

Proceedings

Meeting No 10 "Vibration Analysis Systems" AD 721 912
January 21-22, 1970

Meeting No 11 "Failure Mechanisms: Fatigue" AD 724 475
April 7-8, I970

Meeting No 12 "Identification and Prevention of Mechanical Failures
in Internal Combustion Engines" AD 721 913
July 8-9, 1970

Meeting No 13 "Standards as a Design Tool in Surface Specification
for Mechanical Components and Structures" AD 724 637
October 19-20, 1970

Meeting No 14 "Advances in Decision-Making Processes in Detection,
Diagnosis and Prognosis" AD 721 355
January 25-26, 1971

Meeting No 15 "Failure Mechanisms: Corrosion" AD 725 200
April 14-15, 1971

Meeting No 16 "Mechanical Failure Prevention Through Lubricating Oil Analysis" AD 738 855
November 2-4, 1971

Meeting No 17 "Effects of Environment Upon Mechanical Failures, Mechanisms and Detection" AD 750 411
April 25-27, 1972

Meeting No 18 "Detection, Diagnosis and Prognosis" AD 772 082
November 8-10, 1972

Meeting No 19 "The Role of Cavitation in Mechanical Failures" (NBS SP 394)
October 31-November 2, 1973 COM-74-50523

Note: If PB numbers are not listed, NTIS can identify proceedings by NBS SP number.

Proceedings

Meeting No 20 "Mechanical Failure - Definition of the Problem" (NBS SP 423)
May 8-10, 1974

Meeting No 21 "Success by Design: Progress Through Failure Analysis" (NBS SP 433)
November 7-8, 1974 PB 256 770

Meeting No 22 "Detection, Diagnosis and Prognosis" (NBS SP 436) PB 248 254
April 23-25, 1975

Meeting No 23 "The Role of Coatings in the Prevention of Mechanical Failure" (NBS SP 452)
October 29-31, 1975 PB 257 422

Meeting No 24 "Prevention of Failures in Coal Conversion Systems" (NBS SP 468)
April 21-24, 1976 PB 265 552

Meeting No 25 "Engineering Design" (NBS SP 487) PB 271 743
November 3-5, 1976

Meeting No 26 "Detection, Diagnosis and Prognosis" (NBS SP 494) PB 272 848
May 17-19, 1977

Meeting No 27 "Product Durability and Life" (NBS SP 514) PB 281 878
November 1-3, 1977

Meeting No 28 "Detection, Diagnosis and Prognosis" (NBS SP 547) PB 297 399
November 28-30, 1978

Meeting No 29 "Advanced Composites" (NBS SP 563)
May 23-25, 1979

Meeting No 30 "Joint Conference on Measurements and Standards for Recycled Oil/Systems Performance and Durability" (NBS SP 584)
October 23-26, 1979

Meeting No 31 "Failure Prevention in Ground Transportation Systems" (NBS SP 621)
April 22-24, 1980

Meeting No 32 "Detection, Diagnosis and Prognosis: Contribution to the Energy Challenge" (NBS SP 622)
October 7-9, 1980

Meeting No 33 "Innovation for Maintenance Technology Improvements" (NBS SP 640)
April 21-23, 1981 (not available from NTIS)

Meeting No 34 "Damage Prevention in the Transportation Environment" (NBS SP 652)
October 21-23, 1981

The following MFPG publications were printed by the **Cambridge University Press.**

Proceedings

Meeting No 35 "Time Dependent Failure Mechanisms and Assessment Methodologies"
April 20-22,1982

Meeting No 36 "Technology Advances in Engineering and Their Impact on Detection, Diagnosis and Prognosis Methods"
December 6-10, 1982

Meeting No 37 "Mechanical Properties, Performance and Failure Modes of Coatings"
May 10-12, 1983

Meeting No 38 The Proceedings was not published because of the format of the meeting.

Meeting No 39 "Failure Mechanisms in High Performance Materials"
May 1-3, 1984

Meeting No 40 "Use of New Technology to Improve Mechanical Readiness, Reliability and Maintainability"
April 16-18, 1985

Meeting No 41 "Detection, Diagnosis and Prognosis of Rotating Machinery to Improve Reliability, Maintainability, and Readiness Through the Application of New and Innovative Techniques"
October 28-30, 1986

Meeting No 42 The Proceedings was not published. Inquire at the Vibration Institute regarding the purchase of copies of individual papers.

Meeting No. 43 "Advanced Technology in Failure Prevention"
October 3-6, 1988

The following Proceedings were published by the **Vibration Institute** and the **MFPT Society.**

MFPG Proceedings

Meeting No 44 "Current Practices and Trends in Mechanical Failure Prevention"
April 3-5, 1990

Meeting No. 45 "Focus on Mechanical Failures: Mechanisms and Detection"
April 9-11, 1991

Meeting No. 46 "Economic Implications of Mechanical Failure Prevention"
April 7-9, 1992 **(No longer available.)**

Meeting No. 47 "The Systems Engineering Approach to Mechanical Failure Prevention"
April 13-15, 1993

Meeting No. 48 "Advanced Materials and Process Technology for Mechanical Failure Prevention"
April 19-21, 1994

MFPT Proceedings

Meeting No. 49 "Life Extension of Aging Machinery and Structures"
April 18-20, 1995

Meeting No. 50 "Technology Showcase - Integrated Monitoring, Diagnostics & Failure Prevention" (A Joint Conference)
April 22-26, 1996

Meeting No. 51 "A Critical Link: Diagnosis to Prognosis" (A Joint Conference)
April 15-17, 1997

Available Proceedings may be purchased from

MFPT Society Headquarters
4193 Sudley Road
Haymarket, VA 20169-2420

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 1997	Final Conference Proceedings-April 14-18,1997

4. TITLE AND SUBTITLE
A Critical Link: Diagnosis to Prognosis

5. FUNDING NUMBERS
G: N00014-97-1-0752

6. AUTHOR(S)
Henry C. Pusey and Sallie C. Pusey, Eds.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Society for Machinery Failure Prevention Technology (MFPT)
4193 Sudley Road
Haymarket, VA 20169-2420

8. PERFORMING ORGANIZATION REPORT NUMBER
MFPT 51

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Program Officer Thomas McKenna
Ballston Tower One
800 N. Quincy St., Arlington, VA 22217-5660

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Prepared in cooperation with the Reliability, Stress Analysis and Failure Prevention(RSAFP) Committee of ASME.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
This report is the proceedings of a joint conference, the 51st Meeting of the MFPT Society & the 12th Biennial Conferenceof the RSAFP Com. of the ASME. Approximately 80 technical papers are included on the following topics:Failure Analysis; Diagnostics&Prognostics;Design, Failure Prevention & Stress Analysis; Sensors Technology; Probabilistic Design; Failure Analysis & Safety Methods; Oil Analysis; Signal Processing; Numerical & Analytical Methods; Data Collection; Life Extension & Durability; and Fault Detection & Diagnostics of Vibration Structures -Method & Experiments.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Failure Analysis, failure prevention, diagnostics, prognostics, reliability, stress analysis, sensors, signal analysis, oil analysis	796
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI-Std Z39-18